HEC Montreal

Department of International Business

EXPLORING UNCOVERED BUSINESS OPPORTUNITIES A DATA SCIENCE PERSPECTIVE INTO AMBIDEXTERITY

By

RAMI OSMAN

in partial fulfillment of the requirements
for the degree of Master of Science
in International Business

August 2023

Supervised by Prof. Thierry Warin

Contents

1 Acknowledgements

This dissertation is so dear to me, not just because of its huge potential, but more because of the journey it took me through and the people involved in that journey. It also serves as a reminder whenever I reread this of what I can always achieve, regardless of any challenges facing me. Despite being tough, this journey was very rewarding and humbling after all. And I would not change one bit of it if I had the choice.

To begin with, I would like to thank every single person I've interacted with at HEC Montreal. This includes, but is not limited to, the registrar's office who were always available to answer the difficult questions, student services who supported me throughout, career services who took our post-grad success personally, the janitors and security guards who kept us safe 24/7/365 during COVID and beyond, all the professors who constantly fed my curiosity and helped me push myself out of my comfort zone, in addition to all my classmates who've made the effort to engage in class discussions making the learning process even more valuable. Above all, it is imperative to thank the researchers, scholars and open-source developers whose contributions I've built upon.

During that journey, I've also had a demanding fulltime job that required me to be in different Canadian provinces and US states to keep our economy running. Considering that, I would like to thank the people who believed in me and gave me unique opportunities at work to prove my added value with tangible and intangible results. What I learned from everyone below me and above me in the ranks was invaluable for my growth and progression. I would also like to thank the many people who challenged me at every corner, you gave me an additional reason to prove me right.

Last but not least, a HUGE thank you to Professors Thierry Warin and Nathalie de Marcellis-Warin who have always supported me unconditionally, involved me in many different affairs that helped me grow, created opportunities for me to expand my horizons and influenced me in many different ways to discover my limits and push them further. As for my parents, giving you a Mont-Tremblant covered with gold is still insufficient to thank you enough for an infinite amount of reasons.

It is true that not all heroes wear capes and not all real influencers are on social media.

Finally, I am really excited to experience the next journey of surprises, acquaintances, growth and challenges.

2 Introduction

In 1620, Sir Francis Bacon articulated a seminal observation in his work, Novum Organum: "Human knowledge and human power meet in one; for where the cause is not known the effect cannot be produced" (Bacon 1620). Transitioning to the contemporary era, the same principle remains evident in the corporate landscape. For instance, Kodak's failure to listen to their customers and subsequently their lack of knowledge in adopting novel business models led to a significant reduction in their market power (Anthony 2016). On the other hand, Fujifilm, a principal competitor of Kodak, astutely perceived the impending digital transformation. And through rigorous organisational restructuring, they adeptly maneuvered the complexities of the digital paradigm shift. Furthermore, Fujifilm's strategic alignment of their core competencies with emergent market demands facilitated their pivot into the "anti-aging" cosmetics sector in addition to dominating the tac film technology, an essential component of LCD displays (Fujii 2016).

Other illustrative examples of successful corporate pivots encompass Netflix's evolution from a DVD distribution model to a globally renowned streaming service ("Netflix Reed Hastings" 2017), Play-Doh's metamorphosis from a coal soot cleaning agent to a popular artistic modelling compound (Davis 2011; Klara 2016; Sutton 2002), Western Union's transition from telegram provision to wire transfer services (Eaves 2006; "Western Union Stops Telegrams" 2006), and Honda's astute market analysis that led to the inception of the dirt bike segment in the US (Glaveski 2021). That being said, what are the common factors among all these stories?

A meticulous examination of the aforementioned success narratives identifies four overarching commonalities:

1. An early detection of potential disruptions or opportunities.
2. A comprehensive analysis of these precursors to discern their implications.
3. The extraction of actionable insights from the analysed data.
4. The audacity to implement transformative changes to their business models.

This leads to an imperative inquiry: what mechanisms can organisations employ to detect such subtle market shifts at the outset?

As a matter of fact, the contemporary era is characterized by unprecedented interconnectedness that is largely attributable to the digital revolution. This interconnectivity

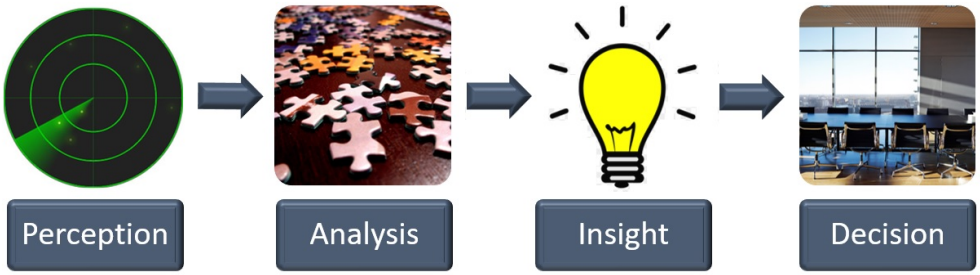


Figure 1: Breakdown of steps taken by companies to pivot successfully

fosters intricate linkages among pivotal business stakeholders, namely: customers, competitors, and shareholders. Online Social Networks (OSNs), exemplified by platforms like Twitter and Facebook, epitomize this trend (Warin 2022). Such networks facilitate the dissemination of pertinent information among these stakeholders. When meticulously analyzed, this information can yield profound insights. Add to that, daily global data generation is estimated to reach 463 exabytes by 2025 (Desjardins 2019), which is equivalent to 463 million terabytes or 463 billion gigabytes. Hence, the prevailing ambiguity in the business milieu is not necessarily a consequence of data shortage but rather an outcome of data inundation, amplifying the complexity of the competitive arena. Therefore, the proficiency to amass, sift through, and interpret this voluminous data becomes crucial.

In essence, contemporary organizational survival mandates the successful navigation of uncertainties to keep up with the rapid changes that are driven by leaps in advancement of technology, an immense degree of connectivity and rapidly growing levels of complexity. The paragons in this environment are perceived to be ambidextrous trailblazers with acute foresight, innovative thinkers transcending conventional paradigms, and astute market players capitalizing on pertinent data. And despite the abundance of frameworks and strategies propounded by eminent academicians, the nexus between organizational ambidexterity and data science remains relatively uncharted, thereby underpinning the impetus for this dissertation.

Data Science is conceptualized as the synergistic amalgamation of statistics, programming, and domain expertise, employed to distill insights from both structured and unstructured data sets (IBM n.d.; Warin and CIRANO 2020). Organisational Ambidexterity, herein referred to as Ambidexterity, is delineated as the capability to harmoniously engage in both explorative and exploitative endeavors, given finite resources (C. A. I. O'Reilly and Tushman 2004). While explorative activities encompass facets like innovation, discovery, and flexibility (March 1991), exploitative activities gravitate towards efficiency, productivity, and execution (March 1991).

This thesis endeavors to operationalize the notion of ambidexterity using data science, with an emphasis on the inherently uncertain and resource-intensive explorative facets.

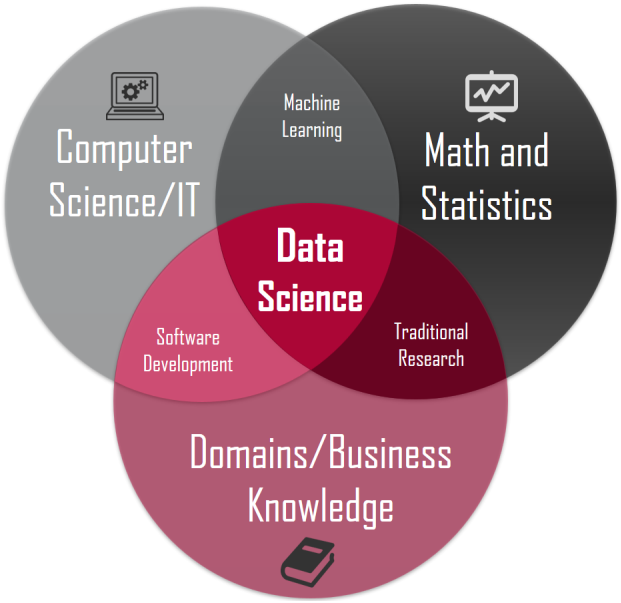


Figure 2: Data Science is the intersection between three disciplines: Math and Statistics, IT in addition to Domain Knowledge. Warin, T. & CIRANO. (2020). MATH60033A: Quantitative Methods in International Business Research. https://warin.ca/math60033a/syllabus.html

Yet, transcending academia, our paramount objective is to substantiate the transformative potential of data science when incorporated into strategic deliberations. Specifically, we aim to elucidate how data science can bolster ambidexterity and mitigate uncertainties, thereby enhancing decision-making efficacy. This translates to empowering organizations with a distinct competitive edge by proficiently harnessing intelligence from platforms like Twitter. Then, leveraging advanced machine learning methodologies, we intend to adeptly navigate the vast data terrains, extracting and presenting salient insights to executive leadership. Graphically, our focus is encapsulated within the initial three stages delineated in figure 1.

But first, why tweets? To being with, several published studies have shown that companies

post on social media to achieve some or all of the following three goals: exhibit their latest and greatest, gain legitimacy and maintain a healthy customer service (Beckmann et al. 2006; Kouloukoui et al. 2023; Lundmark, Oh, and Verhaal 2017; Mette Morsing, Schultz, and Nielsen 2008). We are aware that competitors will not tweet the blueprints of their latest innovation for example, but those reputable cited papers imply that companies will still tweet about their latest updates in order to achieve the three goals we mentioned earlier. Second, several studies have proven that tweets have a sizable effect on how the market behaves since it affects stakeholder perception at the micro and macro levels (Ante 2023; Lacka et al. 2022). Another main reason lies in the plethora of information being shared on Twitter. The latest statistics available on the number of tweets per day was found to be more than 500 million in 2014 or close to 347,200 tweets per minute in 2023 (T. Twitter 2014; "User-Generated Internet Content Per Minute 2022" 2023). Now whether any of these tweets are at all useful or not is another matter and depends on who is asking, which paves the way for the following point.

Customers of the shoe industry for example will not find tweets on flowers particularly useful, unless it is valentine's day. That is to say businesses share their latest updates with their curious customers and interested shareholders alike to engage them in a conversation and avoid the unwanted effects of the "out-of-sight out-of-mind" theory. But different stakeholders are intrigued by different topics which explains the varying engagement levels with different tweets. Moreover, businesses and customers are more easily accessible to each other on twitter which facilitates the customer service experience and greatly reduces response delays (Huang et al. 2018; S. Porter 2021; T. Twitter 2017). It is also about being present where the audience is and building a strong brand (S. Porter 2021).

Besides listening to customers and boasting in front of current or potential shareholders, twitter is about observing competitors and also listening to (on) them, knowing what their latest and greatest is, extrapolating what they are going to do based on what they're tweeting, learning from their mistakes (if any) or finding out what works and what doesn't (T. Twitter 2017). Figure 3 demonstrates our simplified version of the modern global market including the three main stakeholders: shareholders, competitors and customers. The links shown in the figure represent the flow of knowledge in our interconnected and globalized world. In reality, links are also present between the different stakeholder groups.

Figure 3: The 3 main stakeholders that affect market dynamics: shareholders, competitors and customers. Shareholders are shown in purple, competitors are shown in red and customers are shown in blue. The choice of colors, locations and quantities of all three stakeholders shown is insignificant

With all that said, our primary objective is to augment the existing body of literature on organisational ambidexterity by demonstrating the pivotal role of data science in aiding organisations to assimilate, structure, sift through, and interpret vast data volumes. This facilitates informed decision-making, even when confronted with partial information, thereby offering a competitive edge over contemporaries. By leveraging data science, we aim to diminish uncertainties and present innovative methodologies for scrutinising the competitive milieu. A quintessential decision, in this context, pertains to discerning the opportune moments and avenues for strategic pivots, thereby unveiling lucrative business prospects. Such informed decisions inherently bolster an organisation's competitive position.

To actualise this, we envisage the development of a proof of concept, wherein our ambidexterity approach will be applied to a randomly selected enterprise within a niche sector. Specifically, we intend to harness the power of data science to monitor customer engagement metrics for the chosen enterprise and its competitors on Twitter, focusing on three primary activities: replies, retweets, and likes. Concurrently, we aim to gauge shareholder engagement by tracking interactions with news articles pertaining to entities within the niche sector. For this purpose, shareholders will be represented by influential Twitter accounts synonymous

with business news dissemination, akin to the niche sector's equivalent of Bloomberg. Crucially, our endeavor extends to monitoring competitors on Twitter, extracting insights from their tweets and engagement metrics to ensure that our selected enterprise remains abreast of industry developments. Broadening our observational scope to encompass Twitter accounts beyond the immediate industry purview can be instrumental in the exploration phase, reminiscent of Fujifilm's strategic shift from photography to the chemical domain, culminating in the inception of ASTALIFT (FUJIFILM 2007).

The data science methodology encompasses the extraction of tweets and associated metadata within our delineated observational ambit, utilizing Twitter's rest API. Subsequent stages involve the application of data science techniques to eliminate redundant information and analyze the residual data based on predefined criteria. The culmination of this process will be the synthesis of our analyzed data into intuitive visual representations, tailored to guide and influence decision-makers. A comprehensive exposition of these stages will be presented in the ensuing Methodology chapter.

Our research endeavors to provide an empirical answer to the overarching question:

- How can data science serve as a catalyst for organizations to adeptly maneuver through markets characterized by heightened levels of uncertainty, complexity, and competitiveness, thereby achieving ambidexterity?

This central query further breaks down into the following sub-questions:

- To what extent can data derived from Twitter enable companies to anticipate and outmaneuver their competitors?
- How can Natural Language Processing (NLP) techniques be harnessed to expedite and enhance the exploration phase?
- Is it feasible for applied network analysis to illuminate previously unidentified business opportunities by visualizing a network of nodes and edges?

The interplay of these research dimensions is encapsulated in the subsequent Venn diagram, which delineates the thematic scope of this dissertation.

In the ensuing chapter, we embark on a rigorous literature review, leveraging data science methodologies to pinpoint seminal works from the global academic community that delve into the intricacies of ambidexterity and its symbiotic relationship with strategic management. We

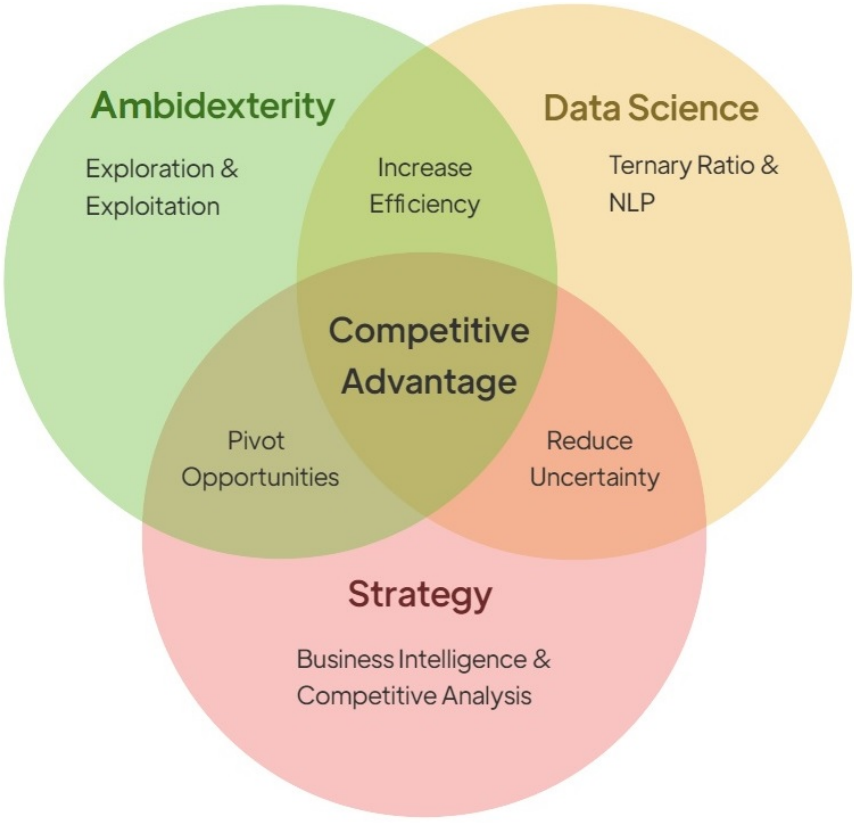


Figure 4: Venn diagram showing the intersections between the main topics of this dissertation

will subsequently delve into a meticulous examination of select papers on organisational ambidexterity. Our objective is to fortify the existing literature by elucidating the transformative potential of data science in enabling organizations to assimilate, structure, and interpret vast data repositories.

Following this, we will explain our methodological approach, detailing the criteria and rationale behind the selection of enterprises within a niche sector, specifically the pulp & paper industry in our study. This section will provide a comprehensive overview of our data acquisition and analytical procedures. Furthermore, we will visually represent our data filtration process, substantiating our methodological choices.

Subsequent sections will be dedicated to the presentation of our empirical findings. We will interpret these results in the context of the broader academic discourse, elaborating on their implications and potential contributions to the field. The discourse will culminate in a contemplation of prospective research trajectories, before presenting our concluding remarks.

3 Literature Review

Our aim in this chapter is to have a rigorous and comprehensive literature review on the topic of organisational ambidexterity, or from hereon referred to as ambidexterity. By fully comprehensive, we are targeting all the published literature on ambidexterity regardless of languages, publication platforms, author locations, affiliations, subjects or publication years among other factors. That not only ensures the comprehensiveness of our literature review, but it also eliminates any innate biases and ensures the equity, diversity and inclusion (EDI) aspect of our process for maximum objectivity. And in order to achieve all that, we performed this literature review by coding in R.

To elaborate, we logged into Web of Science database to search for all documents using the query "Strategy AND (Ambidexterity OR Ambidextrous OR Organisational Ambidexterity OR Organizational Ambidexterity)". We integrated strategy in our search using the AND logic operator since it is one of the core foundations of our research premises to lead us to our goal shown in figure 4 of the introduction chapter: gain a competitive advantage. We also took all the document types available in order to increase our scope of vision on the literature discussing this topic.

As a result, we obtained a total of 1,116 documents. We then refined our search to exclude the following Web of Science categories since they do not pertain to our subject goal presented in the Venn diagram in figure 4: Psychology, Hospitality, Behavioral Sciences, Ethics, Medical and Health Sciences, Political Science, Forestry, Geography, Anthropology, Asian Studies, Chemistry, Criminology, Cultural Studies and Sociology. Even though these exclusions may be perceived as biasing our literature review since those categories might include few pertinent documents (depending on their classification methodology), they only make up a total of 100 documents which is still less than 9% of our 1,116 total results. From a statistical sampling standpoint, this indicates that the exclusions will generally not affect our normal distribution of topics presented. In addition, it would be challenging to build the case that a document from these seemingly far categories could contain information or citations that would radically change the course of our literature review. Furthermore, our goal is to utilize data science for detecting and analyzing high speed signals from OSNs to use in a strategic way, and these exclusions do not appear to add any value that helps us reach our goal. In fact, information from the documents classified within these categories may create noise and distort our analysis.

Besides that, it is important to note that an attempt to include different forms of the word

"data science" within the aforementioned search query led to 0 results on Web of Science. This clearly indicates that our research topic is quite underdeveloped. After excluding the non-pertinent Web of Science categories, we used the full record of the remaining 1,016 pertinent documents to continue our algorithmic literature review.

Proceeding forth, the subsequent section will present the outcomes and insights derived from our algorithmic literature review. This will be complemented by an in-depth exploration of the fundamental concepts elucidated in the key academic works identified.

3.1 Algorithmic Literature Review

The algorithmic meta-analysis of the literature offers a systematic and comprehensive approach whose novelty can be summarized in the following five points:

First, the data science powered review allows us to efficiently identify and filter through the immense amounts of literature that discuss our topic of interest. Compared to a traditional systematic literature review, this proves to be much more rigorous than just choosing a relatively smaller quantity of literature that the maximum human being capacity can identify or accommodate to go through. In other words, we would be fulfilling our goal of comprehensiveness by analyzing the whole population instead of just a sample which brings us to our second reason.

The algorithmic review largely minimizes any conscious or unconscious biases that might influence us to choose certain papers from a select pool of academic publications while eliminating others based on the belief that they add no marginal value to our dissertation. Hence, ensuring the comprehensiveness of our literature review by including all the relevant viewpoints discussing our topic of interest.

Third, the algorithmic literature review streamlines and organizes all the resulting publications in different ways using multiple visuals to help us digest and absorb the contents as well as the key takeaways from the immense amounts of knowledge. More precisely, this novel approach provides a thorough analysis of the three structures of knowledge (K-structures): conceptual, intellectual and social.

Fourth, the algorithmic literature review provides useful metadata on the evolution of the knowledge content of our topic of interest that is otherwise not as clear using the traditional systematic literature review, which leads to our final and most important reason.

The most impactful use of algorithmic literature review is for realtime science mapping and not for measuring science, scientists or science productivity. In other words, this methodology does not only provide structural information on the published literature, its metadata also facilitates content or topic analysis through the identification of the most relevant keywords and their semantic mapping for example.

After that, it is up to the researcher to take extra steps and further drill down into the identified literature and keywords to elaborate upon them and pave the way for answering the research question(s).

3.1.1 Bibliographic Data Quality

The preliminary algorithmic analysis of the 1,016 identified literature shows that the completeness of the bibliographic metadata ranges between good and excellent for all the categories, which clearly indicates the high quality of the bibliographic data we will analyze. Below are the results.

Metadata	Description	Missing Counts	Missing %	Status
AU	Author	0	0.00	Excellent
DT	Document Type	0	0.00	Excellent
SO	Journal	0	0.00	Excellent
LA	Language	0	0.00	Excellent
NR	Number of Cited References	0	0.00	Excellent
WC	Science Categories	0	0.00	Excellent
TI	Title	0	0.00	Excellent
TC	Total Citation	0	0.00	Excellent
CR	Cited References	3	0.30	Good
C1	Affiliation	4	0.39	Good
AB	Abstract	7	0.69	Good
RP	Corresponding Author	15	1.48	Good
ID	Keywords Plus	28	2.76	Good
PY	Publication Year	52	5.12	Good
DI	DOI	61	6.00	Good
DE	Keywords	95	9.35	Good

Figure 5: Bibliographic completeness results of the pubished peer-reviewed documents discussing the interesction between ambidexterity & strategy

3.1.2 Bibliographic Overview

3.1.2.1 Main Information The data presented offers a comprehensive overview of the literature spanning from 1991 to 2023. Over these 32 years, a total of 1,016 documents have been sourced from 352 distinct journals, books, and other academic platforms. This corpus of literature has grown at an annual rate of 14.42%, with the average age of a document being 5.03 years. Each document, on average, has garnered 39.71 citations, culminating in a total of 44,013 references across all documents.

In terms of content, there are 1,657 keywords identified by the indexing service (Keywords Plus) and 2,737 keywords provided by the authors themselves. This suggests a rich diversity of topics and themes covered in the literature.

The literature is the collective effort of 2,481 authors, with 95 of them having penned single-authored documents. Collaboration is a notable feature, with an average of 2.95 co-authors per document. Furthermore, 40.85% of the collaborations are international, indicating a global discourse on the subject.

Breaking down the types of documents, the majority are articles, numbering 853. There are also unique categorizations like articles that are also book chapters (2), articles with early access (49), and articles that are proceedings papers (3). Other document types include editorial materials (2), meeting abstracts (1), proceedings papers (71), and reviews (31), with some of these having early access versions as well.

In summary, the following table provides a holistic view of the academic landscape on the topic, highlighting the depth, diversity, and collaborative nature of the research conducted over three decades.

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan (Years)	1991 to 2023
Sources (Journals, Books, etc)	352
Documents	1016
Annual Growth Rate %	14.42
Document Average Age	5.03
Average citations per doc	39.71
References	44013
DOCUMENT CONTENTS	
Keywords Plus (ID)	1657
Author's Keywords (DE)	2737
AUTHORS	
Authors	2481
Authors of single-authored docs	95
AUTHORS COLLABORATION	
Single-authored docs	102
Co-Authors per Doc	2.95
International co-authorships %	40.85
DOCUMENT TYPES	
article	853
article; book chapter	2
article; early access	49
article; proceedings paper	3
editorial material	2
editorial material; book chapter	1
meeting abstract	1
proceedings paper	71
review	31
review; early access	3

3.1.2.2 Annual Scientific Production The graph depicting annual scientific production serves as a testament to the evolving interest in the subjects of our search query. The initial surge in interest around 2004 indicates the nascent stages of academic exploration into these topics. The subsequent uptick in 2007 suggests a growing recognition of their importance, possibly due to emerging real-world applications or foundational research breakthroughs. The pronounced increase in 2015 underscores a maturation phase, where the topics likely became central to academic discourse, reflecting their significance in contemporary research paradigms. The conspicuous decline post-2022 can be attributed to the temporal limitations of the dataset; the absence of publications beyond this point is a function of the data's currency rather than a waning interest in the subject matter. In essence, the trajectory of this graph underscores the burgeoning relevance of the topics encapsulated in our search query within the academic community over the past two decades.

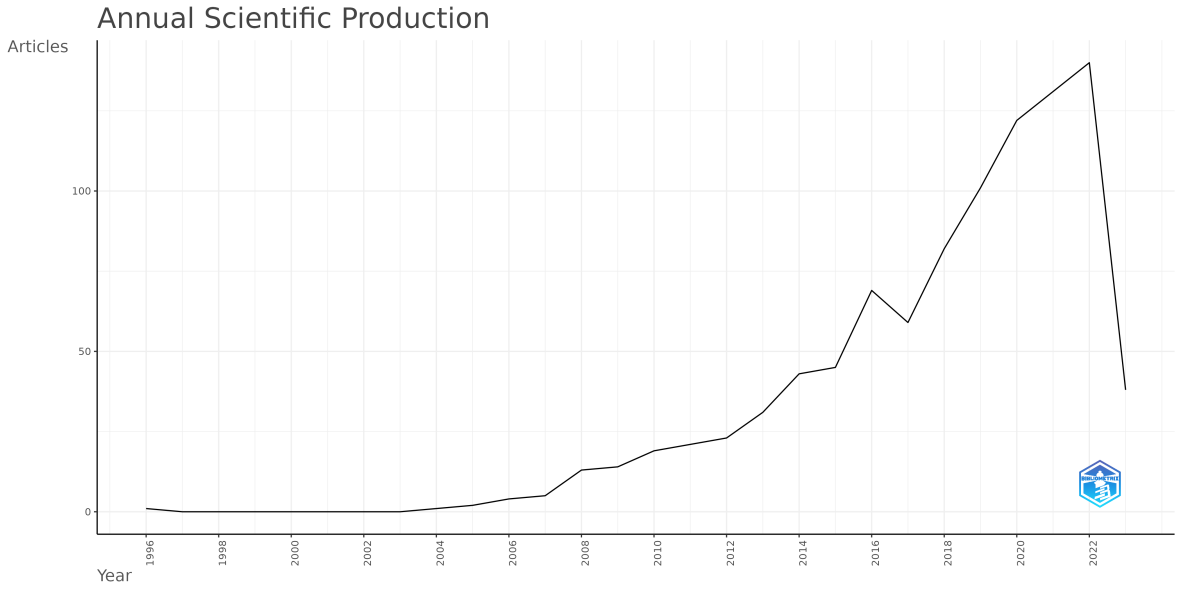


Figure 6: Annual scientific production

3.1.2.3 Three-Field Plot The three-field plot shown on the next page provides a holistic visualization of the interplay between pivotal sources, influential authors, and pertinent keywords within the research domain of interest. By mapping these three dimensions, one can discern patterns of collaboration, influence, and thematic focus.

The leftmost field, showcasing the 20 most cited sources, offers insights into the foundational texts and seminal works that have shaped the discourse in this field. Their connections to specific authors in the central field can indicate direct contributions, collaborations, or influential citations.

The central field, highlighting the 20 most prominent authors, serves as a testament to the key thought leaders and contributors in this domain. The prominence of Michael Tushman, as the leading figure, underscores his seminal contributions and the significant impact of his work on the broader research community. His position and the density of links emanating from him suggest a prolific academic output and a central role in shaping the discourse.

The rightmost field, delineating the 20 most relevant keywords, provides a thematic overview of the prevalent topics and areas of focus within the research domain. The connections between authors and keywords can elucidate the specific areas of expertise or thematic focus of each author.

In essence, this three-field plot not only offers a snapshot of the key players and themes in the research domain but also elucidates the intricate web of relationships, collaborations, and thematic intersections that define the field.

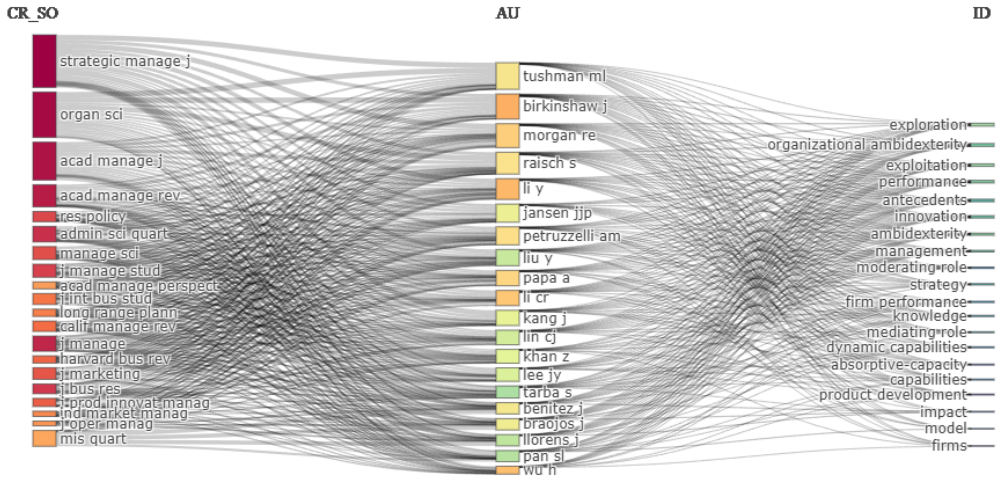


Figure 7: Three field plot

3.1.3 Sources

3.1.3.1 Most Local Cited Sources The subsequent diagram, focusing on the most frequently cited local sources, offers a deeper dive into the foundational journals and publications that have significantly influenced the discourse within the research domain. The prominence of both the "Strategic Management Journal" and "Organisational Science" in this diagram underscores their pivotal role as primary repositories of seminal works and

groundbreaking research in the field.

When juxtaposed with the previous three-field plot, the interconnectedness becomes evident. Leading authors, such as Michael Tushman and Julian Burkinshaw, have either contributed to or frequently cited articles from these esteemed journals, reinforcing their status and influence. Their association with these journals not only highlights the quality and relevance of their research but also indicates the central themes and discussions that are shaping the field.

In essence, the diagram serves as a testament to the symbiotic relationship between influential authors and premier journals. It underscores the iterative nature of academic discourse, where leading authors both contribute to and draw insights from these foundational sources, thereby perpetuating a cycle of knowledge creation and dissemination.

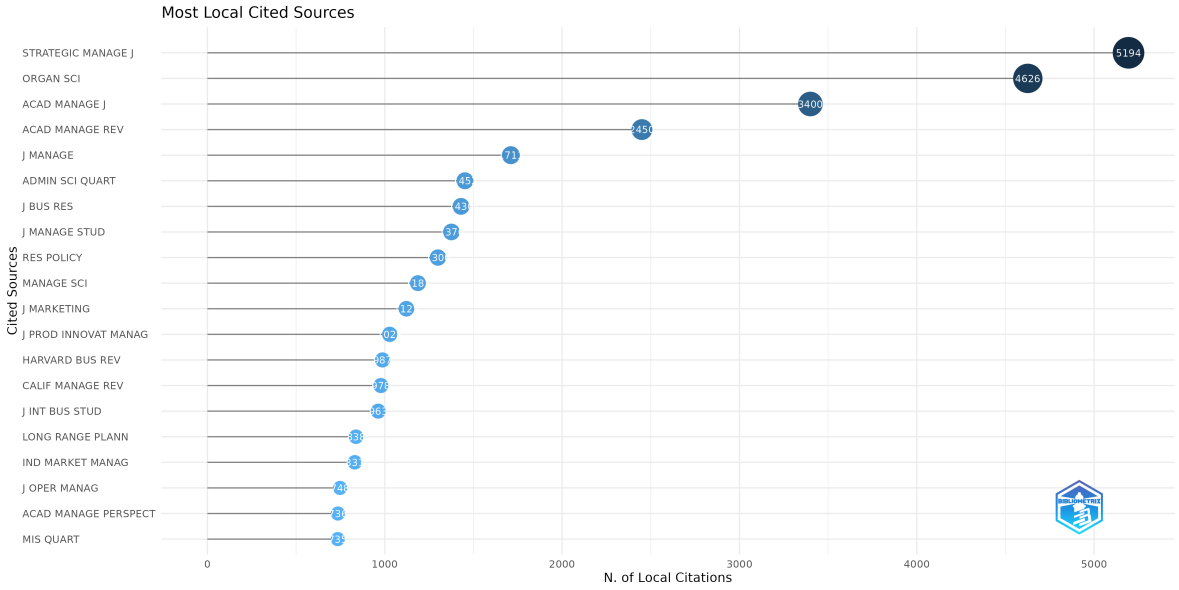


Figure 8: Most local cited sources

3.1.4 Authors

3.1.4.1 Most Local Cited Authors The diagram below shows the 20 most local cited authors. While Michael Tushman, Julian Birkinshaw and Charles O' Reilly dominate the list in the Western Hemisphere, Zi-Lin He and Poh-Kam Wong dominate the list of most local cited authors in the east.

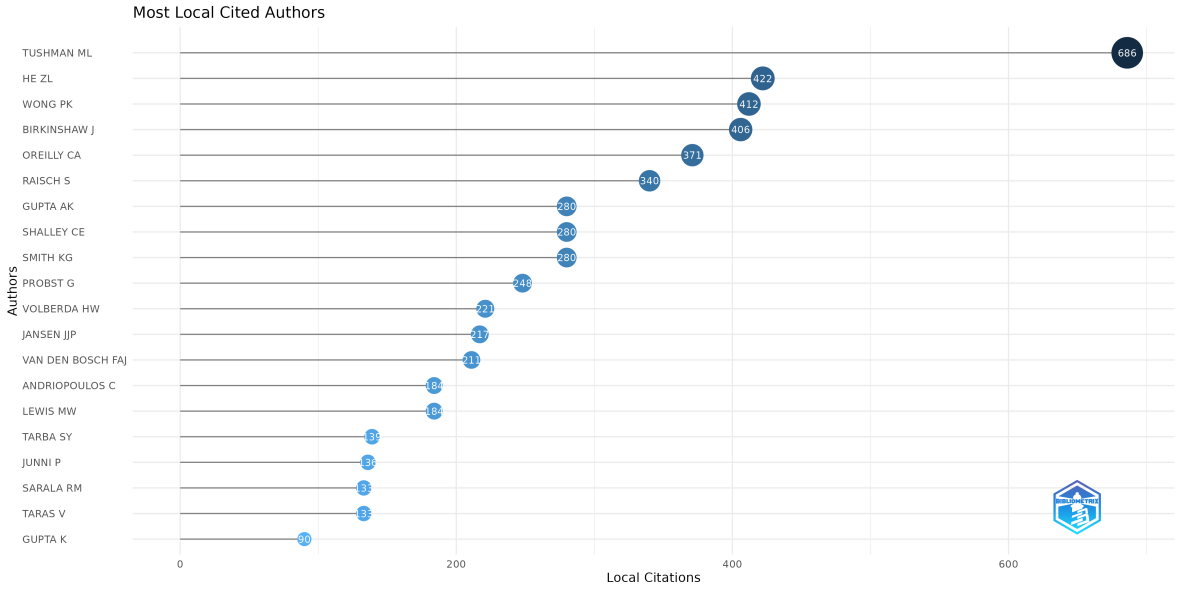


Figure 9: Most local cited authors

3.1.4.2 Most Cited Countries

As shown in the diagram below, the top 3 most cited countries are USA followed by UK followed by China which is coherent with the countries of the most local cited authors seen in figure 9.

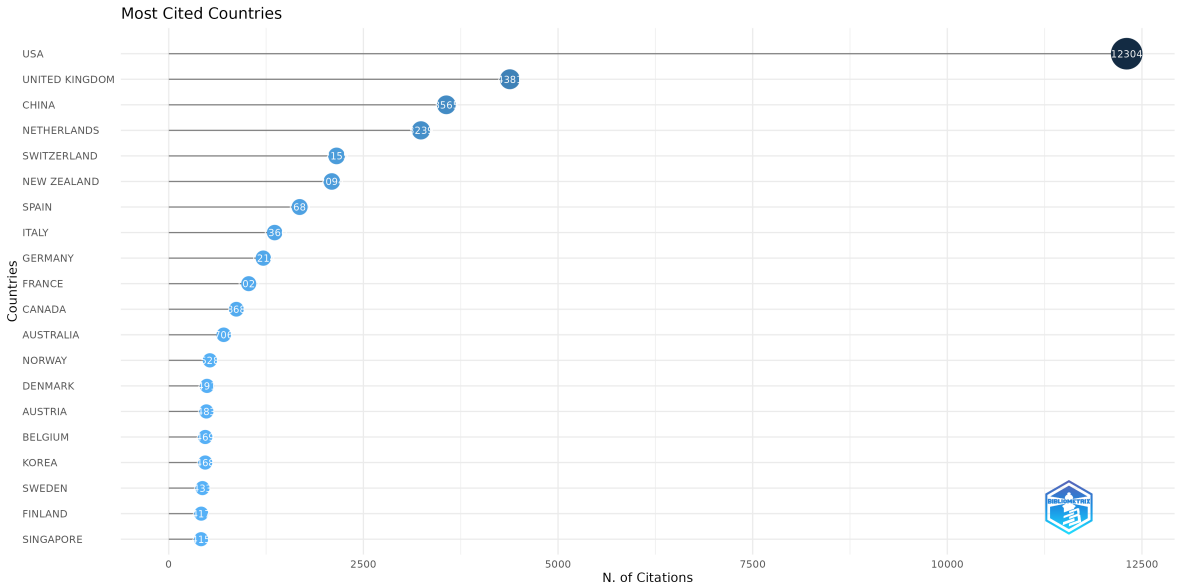


Figure 10: Most cited countries

3.1.5 Documents

3.1.5.1 Most Local Cited References

Figure 11 on the next page shows the most local cited references with the authors specified. Chronologically, it is clear that James March was the spark that ignited the conversation over organisational ambidexterity in 1991, followed by the prominent scholars we have seen such as David Levinthal, Mary Benner, Michael Tushman, Cristina Gibson and Charles O' Reilly from the West. As for the East, Zi-Lin He is the most cited reference.

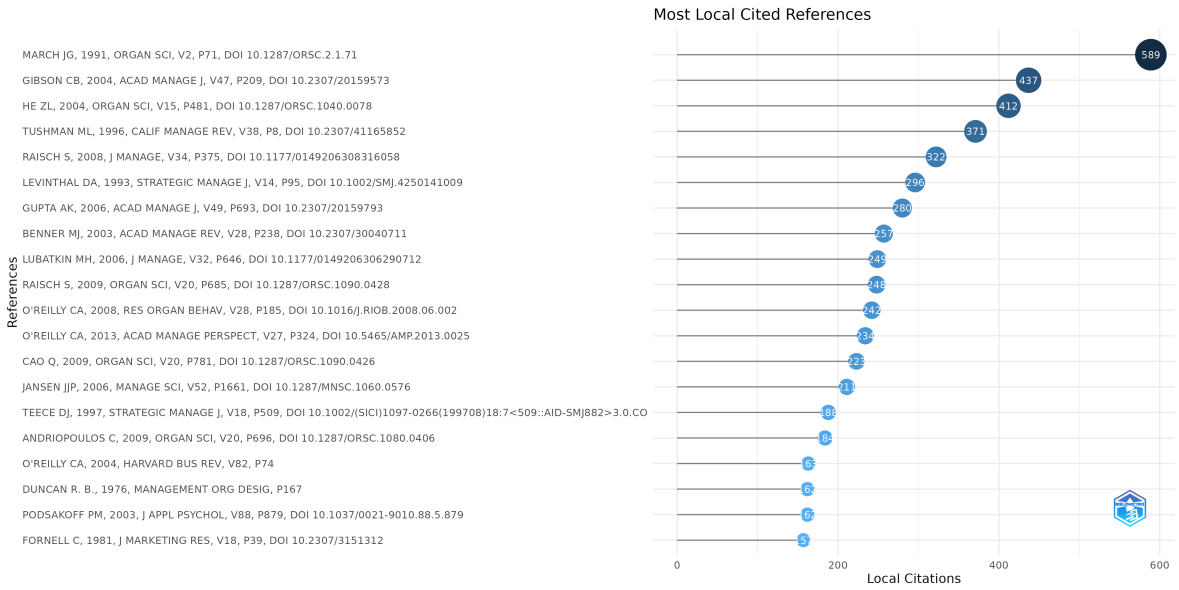


Figure 11: Most local cited reference

3.1.5.2 Most Frequent Words The diagram below shows the 20 most frequent words used within the documents of our search query. Not surprisingly, exploration, exploitation, ambidexterity and performance came in the top ranks. But to further anchor our research premises, strategy and innovation came up in the top 10 ranks which validates the playground of our dissertation. In addition, dynamic capabilities and absorptive capacity along with knowledge also showed up which translates to competitive advantage.

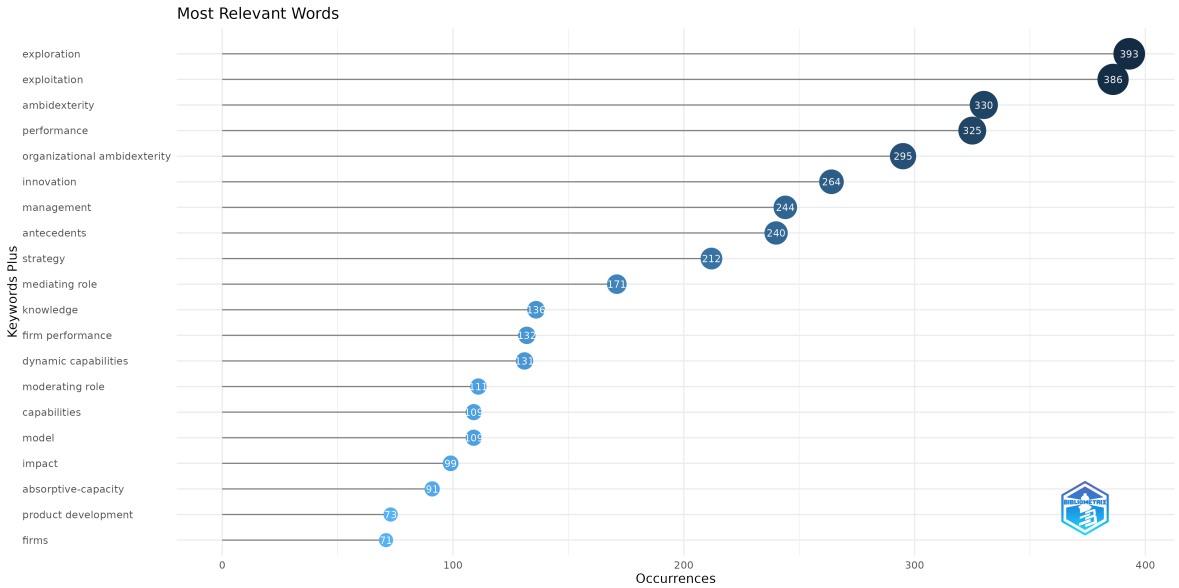


Figure 12: Most frequent words

3.1.5.3 Word Cloud The word cloud below is a different visual representation of the most frequent words.

Figure 13: Most frequent words presented in word cloud format

3.1.5.4 Tree Map The tree map below also adds some color using ratios to the representation of the most frequent words, which is coherent with what we have seen so far in our literature review.

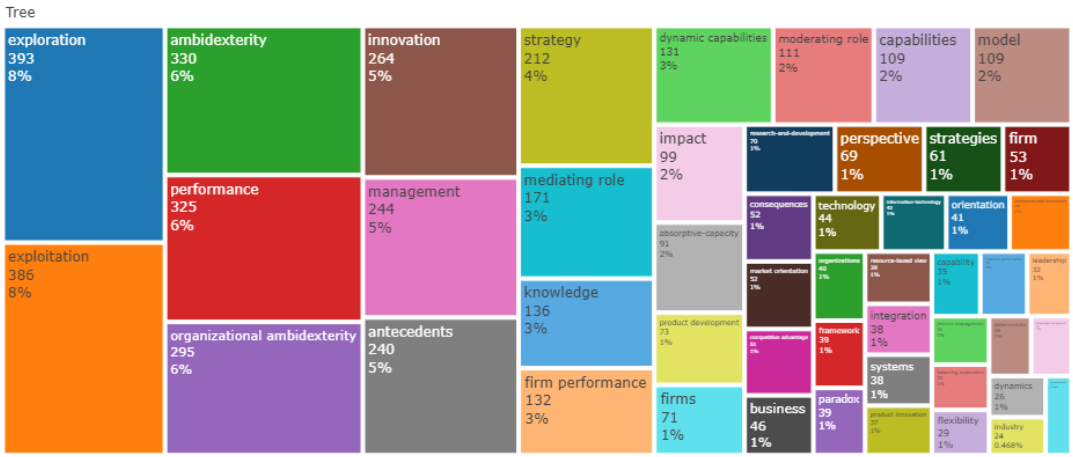


Figure 14: Tree map of most frequent words

3.1.5.5 Word Frequency Over Time Adding a temporal aspect to the word frequency, we can see in the following diagram that exploration, innovation and strategy have started picking up in the last 5 years indicating that these topics are hot within the boundaries of our research relative to the other concepts shown. This is clear evidence that our topic is relevant today more than ever.

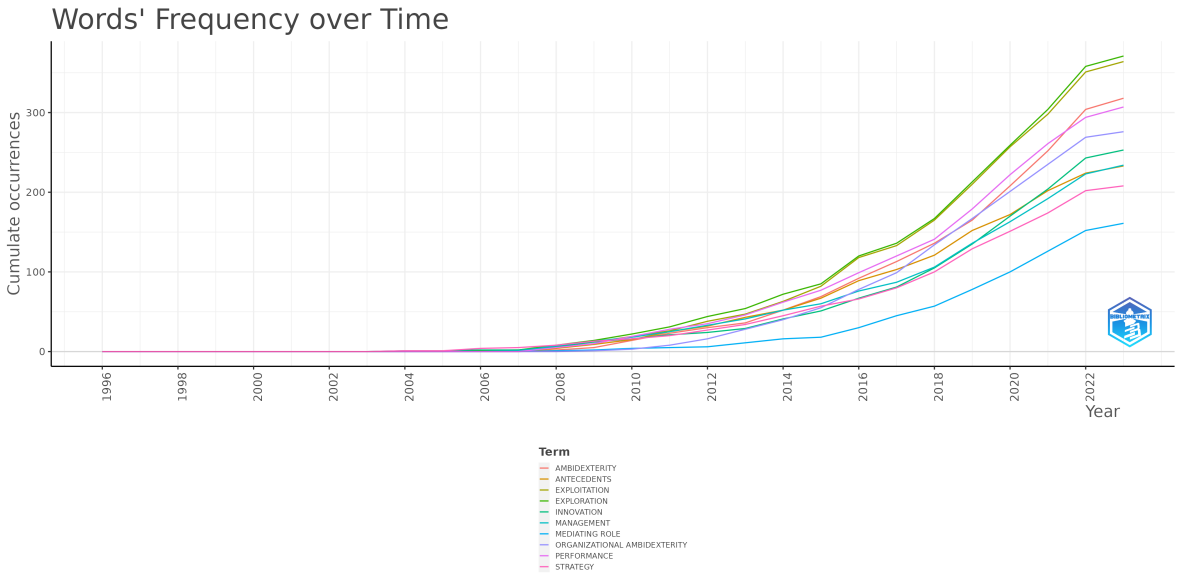


Figure 15: Word frequency over time

3.1.5.6 Trending Topics Based on Keywords The next page shows a diagram of the trending topics based on keywords over time. What is special about this diagram is that it also includes business model innovation as a trending topic with high frequency over the past few years. It is important to note that business model innovation is synonymous to the pivoting success stories discussed in our introduction.

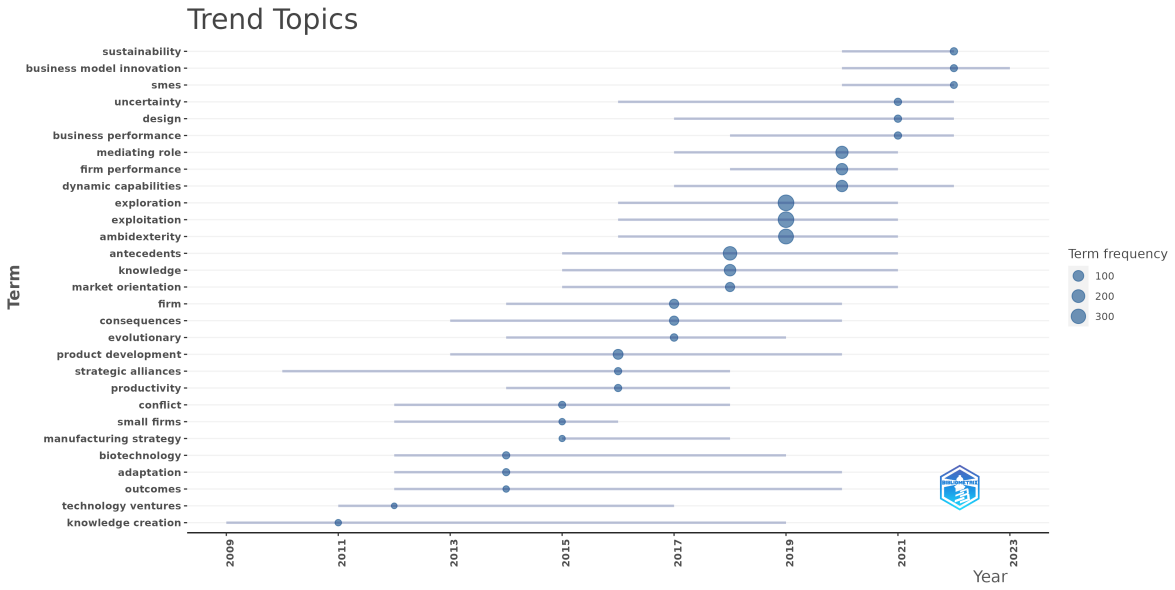


Figure 16: Trending topics based on keywords

3.1.6 Clustering Based on Keywords

3.1.6.1 Network of Authors Related by Keywords The following diagram shows a network map of authors based on mutually cited references at a global level. In other words, if author A cites author C and author B also cites author C, then authors A and B are indirectly related. The colors of the nodes, or dots in the diagram, indicate a cluster that authors belong to based on related keywords. On the other hand, the thickness of the edges, or the lines connecting the nodes, indicate how often this relationship exists between the authors. Another way to represent this relationship is through a digraph.

This graph is particularly important for community detection and mapping of the ego networks which indicates the group of authors that usually work together or that usually cite the same authors (Albert and Barabasi 2002). A good example of this are the authors on the left hand side of the graph, namely: Castillo, Benitez and Braojos. In addition, it is useful to identify small worlds with authors from one cluster embedded within a group of authors in a different cluster. A good example of that is Du Wy in the top center or Volberda in the bottom center. This indicates that these authors are more likely to introduce new or innovative concepts in their communities by building up on their previous publications or citations and collaborating with a new cluster of authors discussing different topics.

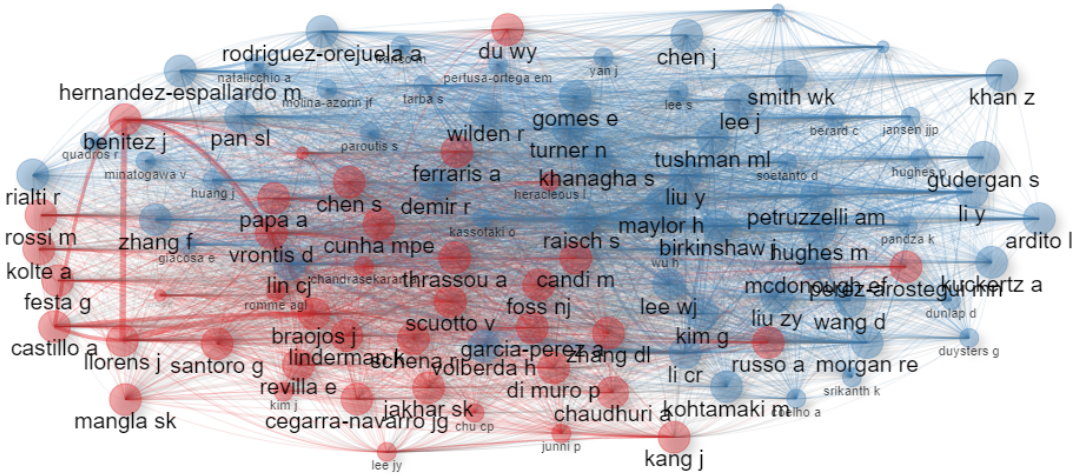


Figure 17: Network of authors based on keywords. The nodes represent the authors and the edges represent two authors that share common keywords. As the size of the node increases, the number of pertinent publications increases. As the thickness of edges increases, the number of common keywords increases.

3.1.7 Conceptual Structure

The conceptual structure dives deeper into the relationships between the different keywords or concepts among the documents from our search query.

3.1.7.1 Co-occurence Network - Visual As can be seen from the following co-occurrence graph, exploration and exploitation are connected by thick edges, or lines. This entails that they appear together in the majority, if not all of the cases. We can also see that the concepts of exploration, exploitation and ambidexterity either appear in the middle of our graph or are connected to the majority of other concepts, which indicates their centrality and importance towards all the other concepts shown. Hence, they form the base to all the other concepts shown.

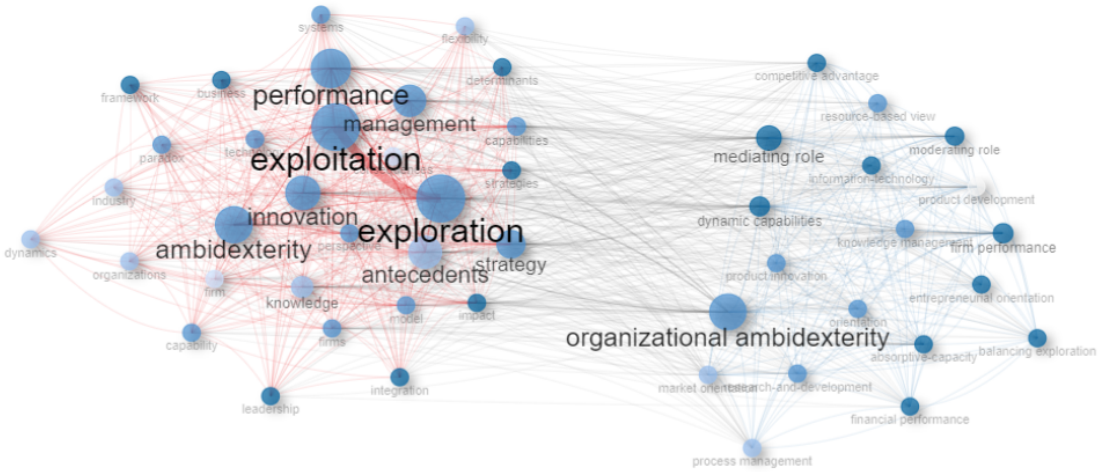


Figure 18: Visual representation of the co-occurance network of concepts. The nodes represent the concepts and the edges represent the co-occurence of both connected concepts within the same publication. As the size of the node increases, the number of occurences of the concept in the resulting publications increases. As the thickness of edges increases, the level of co-occurence in the resulting publications increases.

3.1.7.2 Co-occurence Network - Numeric The subsequent table presents a list of nodes, which are likely keywords or key concepts, and their corresponding betweenness and closeness centrality measures. These measures are commonly used in network analysis to determine the importance and centrality of nodes within a network. Below is a quick overview of what they entail.

1. **Betweenness Centrality**: This metric quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Nodes with high betweenness centrality have significant influence within a network, as they connect disparate parts of the network. In this context, "ambidexterity" has the highest betweenness centrality, indicating that it serves as a pivotal concept bridging various themes or discussions in the literature.

2. **Closeness Centrality**: This metric measures how close a node is to all other nodes in the network. A higher closeness centrality indicates that a node can reach other nodes in the network more quickly. Here, several nodes like "ambidexterity," "exploitation," and "exploration" have similar closeness centrality values, suggesting they are central in the discourse and can quickly connect to other themes or topics.

From the table on the next page, it is evident that "ambidexterity," "exploitation," and "exploration" are central themes, both in terms of bridging different discussions (high

betweenness) and being closely related to other topics (high closeness). Other significant nodes include "performance," "organizational ambidexterity," and "innovation."

On the other end of the spectrum, nodes like "consequences," "market orientation," and "competitive advantage" have lower centrality measures, suggesting they might be more peripheral or specialized topics within the broader discourse.

This table provides a snapshot of the key themes and their relative importance within the research domain, with "ambidexterity" emerging as a central and bridging concept.

Node	**Betweenness**	**Closeness**
ambidexterity	42.0018367711156	0.0204081632653061
exploitation	38.4540911210147	0.0208333333333333
exploration	35.0665723760854	0.0204081632653061
performance	30.7487807864965	0.02
organizational ambidexterity	24.2933012105908	0.02
innovation	18.3885707669696	0.0204081632653061
management	17.9365003914404	0.02
strategy	16.2294916246621	0.0204081632653061
mediating role	8.51485570618881	0.02
firm performance	5.91824586870639	0.0181818181818182
knowledge	5.3809119412694	0.0188679245283019
dynamic capabilities	4.4356346419445	0.0192307692307692
capabilities	3.50419992477698	0.0192307692307692
absorptive-capacity	3.05702266857897	0.0181818181818182
moderating role	2.8534545379413	0.0175438596491228
impact	2.84380052948773	0.0192307692307692
model	2.5472657223046	0.0188679245283019
firms	2.02278801734116	0.0188679245283019
product development	1.44771782128266	0.0175438596491228
strategies	1.1660727306412	0.0175438596491228
research-and-development	1.02369649607762	0.0166666666666667
competitive advantage	0.841891136160521	0.0169491525423729
market orientation	0.83564527839241	0.0161290322580645
consequences	0.44750381698131	0.0163934426229508

3.1.8 Intellectual Structure

The intellectual structure represents the relationships between references or citations.

3.1.8.1 Co-citation Network - Visual As clearly shown below, March's 1991 publication on organisational ambidexterity proves to be the most central, i.e. having the highest betweenness centrality, since it sits in the middle and appears to have the highest degree centrality simultaneously. This indicates that March's publication was the springboard for many other publications on that topic at the local and global levels. It is important to note that the colors of the nodes indicate the different clusters identified by the algorithm as co-citation communities.

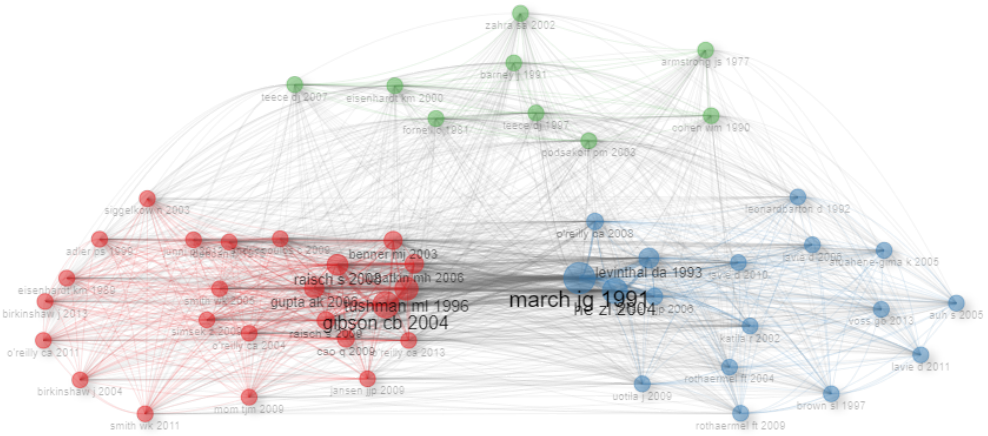


Figure 19: Visual representation of the co-citation network of authors. The nodes represent the authors and the edges represent the co-citation of between both connected authors. As the size of the node increases, the number of co-citations increases. The thickness of the edges has no significance.

3.1.8.2 Co-citation Network - Numeric The results below again prove our synthesis in the previous subsection. This is also coherent with the results we have seen in the previous subsections in the literature review.

Author	Betweenness	Closeness
march jg 1991	104.473989268627	0.0153846153846154
he zl 2004	58.7609096747492	0.0153846153846154
levinthal da 1993	29.014220152513	0.0153846153846154
gibson cb 2004	23.5653144394955	0.0138888888888889
o'reilly ca 2008	23.1348975932063	0.0153846153846154

Author	Betweenness	Closeness
tushman ml 1996	17.6518047045747	0.0138888888888889
jansen jjp 2006	17.0001266450403	0.0153846153846154
gupta ak 2006	13.775637941988	0.0138888888888889
raisch s 2008	12.844638807795	0.0138888888888889
benner mj 2003	11.7302587897073	0.0138888888888889
lubatkin mh 2006	11.1184912365233	0.0138888888888889
cao q 2009	9.08181395998908	0.0138888888888889
lavie d 2010	8.81479507151081	0.0153846153846154
teece dj 1997	7.8161262651722	0.0175438596491228
raisch s 2009	7.68964690328488	0.0138888888888889
podsakoff pm 2003	6.10800886270003	0.0175438596491228
cohen wm 1990	5.23626011965821	0.0172413793103448
eisenhardt km 2000	5.23435209084304	0.0175438596491228
katila r 2002	5.19594987316506	0.0153846153846154
o'reilly ca 2013	5.17399582256088	0.0138888888888889
fornell c 1981	4.47560044314415	0.0175438596491228
simsek z 2009	4.45779110889581	0.0138888888888889
uotila j 2009	4.38034489284103	0.0153846153846154
andriopoulos c 2009	4.19931381952076	0.0138888888888889
lavie d 2006	4.17941387465407	0.0153846153846154
teece dj 2007	4.15313532612733	0.0175438596491228
barney j 1991	3.37998974694008	0.0175438596491228
o'reilly ca 2004	2.95576180203983	0.0138888888888889
smith wk 2005	2.9001440986582	0.0138888888888889
leonardbarton d 1992	2.89560010119842	0.0153846153846154
rothaermel ft 2004	2.87280153660663	0.0153846153846154
duncan r. 1976	2.72992694870306	0.0138888888888889
atuahene-gima k 2005	2.60998479962837	0.0153846153846154
rothaermel ft 2009	2.53199325995115	0.0153846153846154
jansen jjp 2009	2.46315551331816	0.0138888888888889
auh s 2005	2.3318550298562	0.0153846153846154

3.1.8.3 Historiograph The historiograph below is a directed graph with nodes representing publications and edges representing a citation. This shows how the citations propagated from one document to another over time. It can be inferred that Michael Tushman's 1996 publication was a pivotal document for all the other future publications. It can also be inferred that citations were global and not local as can be seen with the directed graph between Michael Tushman's 1996 publication and Zi-Lin He's 2004 publication.

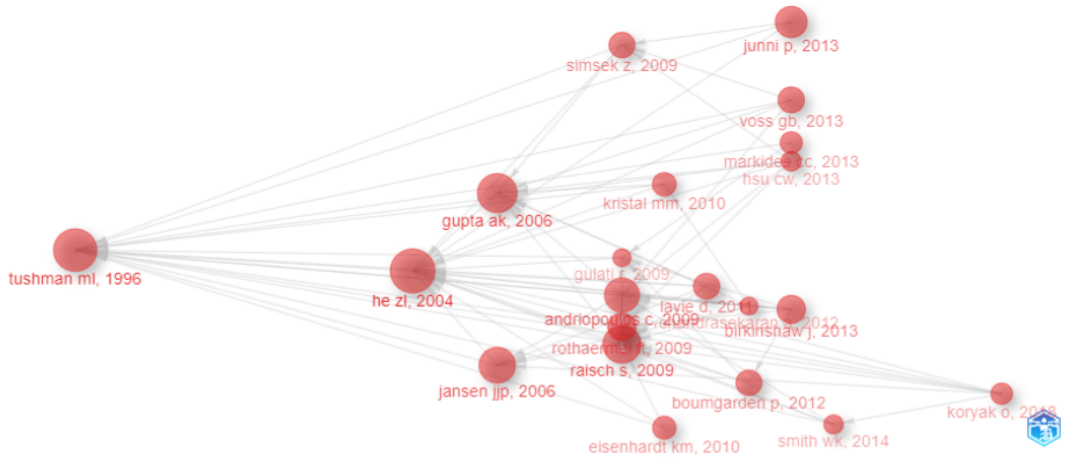


Figure 20: Historiograph of publications and citations on the intersection between the topics of ambidexterity and strategy. The nodes represent the authors and the edges represent the citation from/to the connected authors. As the size of the node increases, the number of citations increases. The thickness of the edges has no significance.

3.1.9 Social Structure

The social structure depicts how authors from different but related disciplines work on topics. Applied Network Analysis on the Social Structure can highlight small worlds and author communities.

3.1.9.1 Collaboration - Network As clearly depicted in the following graph, we can see the different communities of collaborating authors at the local level.

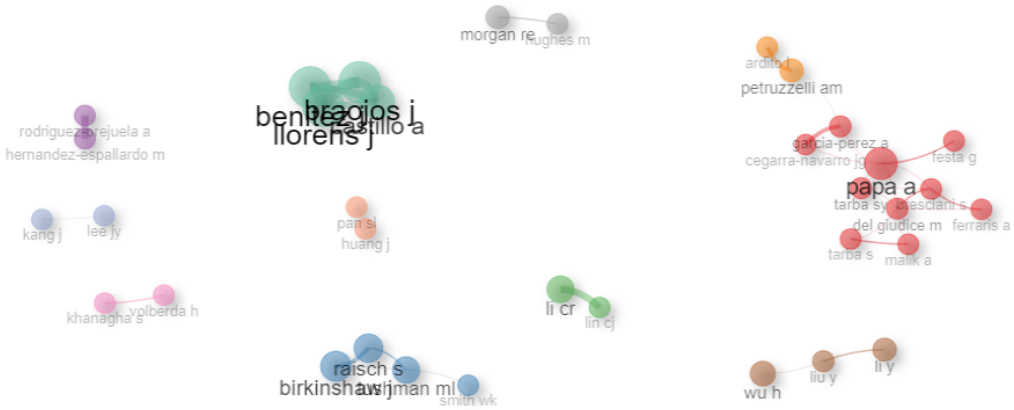


Figure 21: Collaboration networks of authors in their localised communities. The nodes represent the authors and the edges represent the level of collaboration between the connected authors. As the size of the node increases, the number of pertinent publications increases. As the thickness of the edges increases, the level of collaboration increases.

3.1.10 Summary of Algorithmic Literature Review

Since the number of publications is increasing at a fast pace making it difficult to keep up with all the different academic contributions, and considering the importance of understanding and summarizing past research to progress the continuity of knowledge, the algorithmic literature review shows huge potential to fill that gap. Indeed, this methodology proved its ability to provide structured analysis to a massive amount of knowledge, deduce changes in trends and themes over time, identify the shifts in the boundaries of academic disciplines as well as identify prominent academicians and sources. But most importantly, the algorithmic literature review provides a clear and concise big picture of the large body of research discussing the topic(s) of interest.

As a result, we acquired 1,016 documents relevant to our search query "Strategy AND (Ambidexterity OR Ambidextrous OR Organisational Ambidexterity OR Organizational Ambidexterity)". Our analysis showed that the concept of our search query is relatively new with an average age of approximately 5 years, growing at an annual rate of around 14% and an average of roughly 40 citations per document. The concept proved to be highly versatile with more than 1,500 keywords outlining all these publications, of which 41% are a result of collaborations among different authors globally on prominent publication platforms such as the Strategic Management Journal, Organisational Science and Academic Management Journal.

While James March sparked the conversation about ambidexterity in 1991 (March 1991), Michael Tushman built on that position through numerous collaborations to produce several pivotal publications including his co-authored paper with Charles O'Reilly in 1996 (Michael L. Tushman and Charles A. O'Reilly 1996) in addition to his award winning publication with Mary Benner in 2013 (Michael L. Tushman and O'Reilly 2013).

In terms of content, the most relevant keywords form a dense network summarizing the core concepts of our search query. These keywords include ambidexterity, exploration, exploitation, strategy, performance, innovation, business models, knowledge, dynamic capabilities and absorptive capacity. Naturally, ambidexterity has the overarching presence followed by exploration and exploitation. But taking a closer look, it is clear that exploration takes the lead in terms of frequency over time and the rate of increase of that frequency which stipulates three important points: First, our research premises is relevant today more than ever. Second, exploration is the real driver of the train of the actionable keywords such as

strategy, innovation and performance. Third, there are still vast opportunities to develop the implementation of the exploration process when compared to exploitation. Hence, providing a stronger motivation to pursue our research premises shown in figure 4 in the introduction chapter.

Most notably, the diagrams highlight how all the relevant keywords are derived from both exploration and exploitation which are in turn derived from ambidexterity. Besides, business model innovation, which translates to the pivoting examples presented in the introduction chapter, has been trending since the year 2020 with increasing frequency which provides an irrefutable evidence of how exploration strongly accords with strategy and pivoting.

Following our algorithmic literature review using code in R, and after proving its benefits to identify the core concepts, key authors and pivotal publications related to our dissertation, what follows is a deeper dive into how this literature described ambidexterity, its activities, benefits and challenges.

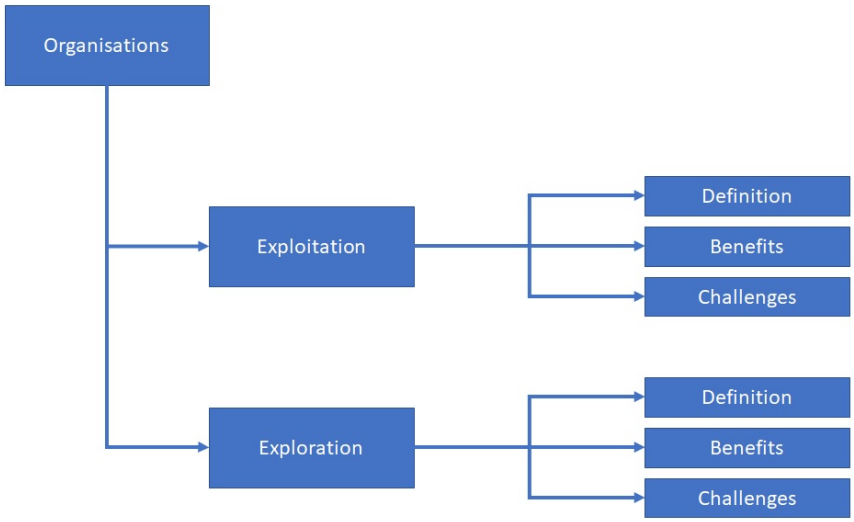


Figure 22: Plan map of the elaboration on the identified literature

3.2 Elaboration on the Identified Core Concepts and Pivotal Publications

3.2.1 The Concept of Organisational Ambidexterity

Per Merriam-Webster's English dictionary, being "ambidextrous" means a person is able to use both hands or feet with equal ease ("Ambidextrous Definition & Meaning - Merriam-Webster" n.d.). Applied to organisations, it describes those that are successful at balancing both activities in which organisations engage, namely: exploration and exploitation (C. A. I. O'Reilly and Tushman 2004). Explorative activities allow an organisation to be creative and adaptable, whereas exploitative activities are related to routine tasks where efficiency plays a role (March 1991). Explorative activities are synonymous to search, discovery, autonomy, innovation and flexibility (March 1991). On the other hand, exploitative activities lean more towards certainty, variance-reduction, efficiency, productivity and execution (March 1991).

Summing up, organizational ambidexterity is described as the ability of an organization to both explore and exploit, i.e. compete in mature technologies and markets where efficiency, control and incremental improvements are prized while also competing in new technologies and markets where flexibility, autonomy and experimentation are needed (Benner and Tushman 2015). By now, one of the main barriers to becoming ambidextrous should be clear - lack of sufficient resources to simultaneously explore and exploit. Evidently, organisational ambidexterity was characterized as the trade-off between efficiency and flexibility, or in other words, the paradox of administration (Thompson 2003). Moreover, March noted that the fundamental adaptive challenge facing firms was the need to both exploit existing assets and capabilities to provide for sufficient exploration and to avoid being rendered irrelevant by the rapid changes in market and technology (March 1991).

3.2.2 The Components, Benefits and Challenges of Ambidexterity

Using another lens, Abernathy, Benner and Tushman suggested that a firm's focus on productivity gains inhibits its flexibility and ability to innovate (Abernathy 1978; Benner and Tushman 2003). Abernathy also discovered that economic decline in the automobile industry during that time was related to seeking efficiency & productivity (Abernathy 1978). One way organisations seek efficiency and productivity is through process management. Process management considers the organization as having a system of interlinked processes that are mapped to improve and adhere to the organization's goals. Its methods affect variation creation through the use of statistical techniques to increase predictability and certainty by

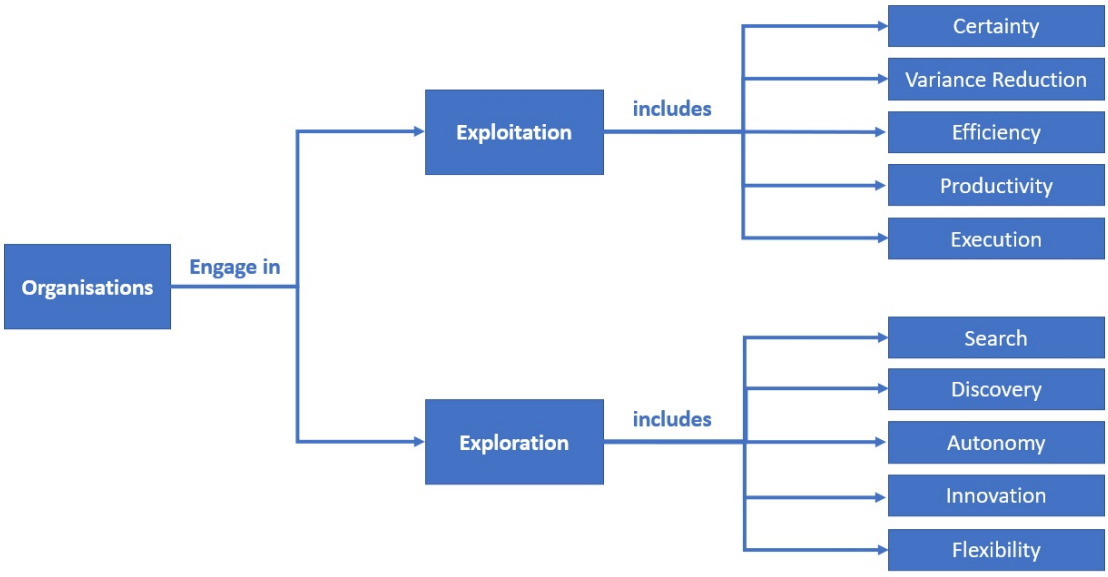


Figure 23: Visual representation of organisational ambidexterity and its components

reducing variation. Empirical results suggested that, in the digital camera industry for example, more extensive process management activities dampened a firm's ability to keep up with rapid tech changes through new product introductions (Benner 2009). Process management is also perceived as steering the organization towards certainty and predictability which encourages its use in other activities throughout the organization, i.e the diffusion of process management to innovative business units in the organization favors exploitative innovation VS. exploratory innovation. As a result, this reduces responsiveness to new markets or new customer segments. Empirical research has shown that process management was not related to long-term benefits and was more related to poor financial performance over longer time (Benner and Tushman 2003). In other words, the spread of process management techniques beyond manufacturing gives rise to unintended effects on innovation and adaptation where the core dynamic capabilities become core rigidities or competency traps in rapidly changing environments (Benner and Tushman 2003). That being said, it should be clear now that the second main barrier for organizations to become ambidextrous is that there is a bias in favor of exploitation with its greater certainty of short-term success. Add to that, exploration by nature is inefficient and is associated with an unavoidable increase in the number of bad ideas. But without exploration, firms will likely fail in the face of change (March 1991; Michael L. Tushman and O'Reilly 2013). Moreover, Tushman added that organizational ambidexterity can be achieved through studying absorptive capacity, dynamic

capabilities and organizational learning in order to ensure it permeates well in the organization (Michael L. Tushman and O'Reilly 2013). For that reason, involving data science with the ambidexterity process largely minimizes the unwanted effects that accompany the exploration process while reaping at least the same amount of benefits. But how are dynamic capabilities related to ambidexterity?

Dynamic capabilities are defined as the firm's ability to integrate, build and reconfigure internal and external competencies to address rapidly changing environments (Michael L. Tushman and O'Reilly 2013). Hence, they are a key part of organizational ambidexterity. In multiple highly cited literature, it was stated that a firm's ability to compete over time may lie in its ability to both exploit and explore (Michael L. Tushman and O'Reilly 2013; P. Ghemawat and Ricart Costa 1993; March 1991; Weick 1979). In other words, a firm's ability to compete over time may lie in its ability both to integrate and build upon its current competencies while simultaneously developing fundamentally new capabilities (Teece, Pisano, and Shuen 1997).

But going back to one of the main barriers, it has been proven that the basic problem an organization faces in our modern times is the balance between exploitative activities to ensure its current viability while at the same time devoting enough energy and resources to exploration to ensure its future viability (C. A. O'Reilly and Tushman 2011). Drilling into this particular paper, the authors specified that dynamic capabilities underpin the ability of firms to sense, seize and reconfigure organizational assets to adapt to the changing environmental conditions. Consequently, sustained competitive advantage comes from the firm's ability to leverage and reconfigure its existing competencies and assets in ways that are valuable to the customer but difficult for competitors to imitate. Particularly dynamic capabilities are rooted in organizational processes or routines around coordination, learning and transformation. They allow firms to sense opportunities and then seize them by successfully allocating resources, often by adjusting existing competencies or developing new ones to address the emerging threats and opportunities.

Although organizational ambidexterity may, under some conditions, be duplicative and inefficient (Ebben and Johnson 2005), the empirical evidence suggests that under conditions of market and technological uncertainty, it typically has a positive effect on firm performance. That is why we are integrating data science into the process of ambidexterity - to eliminate that uncertainty all year long and reduce the inefficiencies caused by ambidexterity as well as

free up resources that could be concentrated to increase firm performance. The study by Annalies Geerts et al. showed that organizational ambidexterity had a positive effect on firm growth after studying more than 500 firms over a period of four years (Geerts, Blindenbach-Driessen, and Gemmel 2010). This was reaffirmed by Michael Tushman that in uncertain environments, organizational ambidexterity appears to be positively associated with increased firm innovation, better financial performance and higher survival rates (Michael L. Tushman and O'Reilly 2013). In addition, it was proven that sequential ambidexterity may be more useful in stable, slower moving environments (such as in the services industry) and for smaller firms that lack the resources to pursue other more complex types of organizational ambidexterity. As Brown, Levinthal, Nelson and their colleagues said, technological innovation is a central engine of organizational adaptation (Brown and Eisenhardt 1998; Levinthal 1991; Nelson and Winter 1982). Innovation is measured along 2 factors: proximity to the original technology and proximity to the existing market. In terms of proximity to the original technology, we have incremental innovation which considers building upon existing technologies, and we have radical innovation which shows a fundamental change to technological trajectory. Coming to proximity to existing markets, there are the exploitative activities which are concerned with existing customers and the exploratory activities which are more concerned with new markets or customer segments since they require new knowledge or departures from existing skills (Levinthal and March 1993; March 1991). The explanation of ambidextrous organization idea and the importance of ambidexterity is contrasted through different cases of companies where innovations are crucial for competitive advantages, such as in the technology sector (Michael L. Tushman and O'Reilly 1999) or for instance, in the pulp, packaging or pharmaceutical industries (Maeir 2015). Subsequently, an organization that lacks exploration in one period may be excluded from areas of future exploratory activities because it has no relevant knowledge base (Cohen and Levinthal 1990). Hence, exploratory activities would decrease and the absorptive capacity of a firm is negatively affected. To counter those unwanted effects, ambidextrous senior teams must develop processes for establishing new forward-looking cognitive models for exploration units while allowing backward-looking experimental learning to unfold for exploitative units.

Revisiting their award winning article after 10 years, Benner et al. mentioned that the increased modularization of products and services along with the simultaneous sharp decline in communication and computation costs due to the digital revolution actually pushed the locus of innovation beyond the boundaries of the firm to open or peer communities (Benner

and Tushman 2015; Adner 2002; Afuah and Tucci 2013; Chesbrough 2006). To complement that, ambidextrous organisational designs require high differentiation, targeted structural integration at points of leverage between exploitation and exploration in addition to senior team integration. And this proves to be another challenge deterring firms from venturing into exploration - the enemy of exploitation. It was also proven that information processing, storage and communication costs in addition to intellectual property challenges have been constraints on innovation (Benner and Tushman 2015). This entails that there is strategic value in organizations that were able to deal with and successfully manage the paradox of exploration VS. exploitation. This proves to be another area where integrating data science into the process of ambidexterity becomes useful: knowing how competitors are operating their businesses to identify their strengths and weaknesses in addition to the threats and opportunities in the industry in order to create winning strategies for sustained competitive advantage.

Add to that, Lis et al.'s systematic literature review on ambidextrous organizations underscores a fundamental tenet of modern management: the inevitability of change. In today's dynamic business landscape, change is not just frequent but is the very fabric of organizational existence. This perspective posits that organizations cannot afford to remain static; they must either react to external shifts or proactively anticipate and drive change to remain relevant and competitive (Lis et al. 2018).

Building on this notion, Luo et al. delve deeper into the characteristics of ambidextrous organizations. They emphasize the unique capability of such organizations to navigate and manage inherent organizational paradoxes. These are situations where organizations face seemingly contradictory objectives, yet, through ambidexterity, they can pursue and achieve both (Luo and Rui 2009). For instance, while the immediate pressures of short-term survival might seem at odds with strategies for long-term growth, ambidextrous organizations can adeptly balance and pursue both. Similarly, they can innovate incrementally while also seeking radical breakthroughs, and they can compete aggressively while also forging cooperative alliances. This duality, as highlighted by Luo and Rui, is at the heart of ambidexterity, allowing organizations to transcend traditional boundaries and achieve multifaceted success (Luo and Rui 2009).

3.2.3 Our Contribution

With all that being said, the essence of this research lies in its novel approach to integrate data science with strategic management, particularly within the realm of organizational ambidexterity. Our distinct contributions are as follows:

1. Strategic Uncertainty Reduction: By weaving data science into the fabric of corporate strategy, this research underscores the pivotal role of data-driven insights in mitigating strategic uncertainties. In an era where decisions are often fraught with risks, especially those associated with exploratory ventures, data science offers a beacon of clarity, helping organizations navigate the murky waters of strategic decision-making.

2. Efficient Exploration: The research posits that possessing comprehensive knowledge of the industry landscape can significantly streamline exploratory endeavors. By harnessing the power of data science, organizations can sieve through vast amounts of information, distilling only the most pertinent insights. This not only economizes resource allocation but also ensures that exploratory ventures are grounded in robust data-driven insights.

3. Unlocking Organizational Potential: At its core, data science is a tool for revelation. By meticulously mapping the business landscape, it can unearth latent linkages between stakeholders, spotlighting innovative pathways and pivot opportunities that might have otherwise remained obscured.

While organizational ambidexterity encompasses both exploration and exploitation, this dissertation hones in on the former, addressing the challenges and barriers highlighted in the comprehensive literature review. Building on the foundational concepts elucidated in the review, the subsequent chapters will delve into the synergistic interplay between data science and strategic management. The aim is to harness high-velocity signals from Online Social Networks (OSNs), transforming them into actionable strategic insights that can confer a competitive edge.

The subsequent table summarizes the pivotal concepts related to organizational ambidexterity, offering readers a concise overview of the foundational themes underpinning this research.

Dimension	Concept	Description	Author(s)
Ambidexterity	Exploitation activities	Exploitation activities are those that are related to routine tasks where efficiency plays a role	March 1991; Benner and Tushman 2015
Ambidexterity	Exploratory activities	Exploration activities are those that allow an organisation to be creative and adaptable	March 1991; Benner and Tushman 2015
Advantages of Ambidexterity	Pivot opportunities	Discovering latent business opportunities and transition to new business models	Lis et al. 2018; Luo and Rui 2009
Advantages of Ambidexterity	Innovation	Non-ambidextrous organisations risk the ability to introduce new products, create new ways of doing business or losing new market opportunities. Innovations is crucial for competitive advantages	O' Reilly 1999

Dimension	Concept	Description	Author(s)
Advantages of Ambidexterity	Competitive advantage	Sustained competitive advantage comes from the firm's ability to leverage and reconfigure its existing competencies and assets in ways that are valuable to the customer but difficult for competitors to imitate, i.e. dynamic capabilities	C. A. O'Reilly and Tushman 2011
Challenges of Ambidexterity	Paradox of Administration	Lack of sufficient resources forces organisations to trade-off between exploratory and exploitation activities	Thompson 2003; O' Reilly & Tushman 2011
Challenges of Ambidexterity	Biases towards certainty	Exploitation is perceived as having greater certainty contrary to Exploration which is perceived as inefficient and is associated with an unavoidable increase in the number of bad ideas	March 1991; Benner & Tushman 2003; Tushman and O'Reilly 2013

4 Methodology

Synthesizing the discussions thus far, the core of this research endeavor is to operationalize and validate a data science-driven approach to organizational ambidexterity, particularly the explorative aspect which is the focus of this dissertation. The chosen method for this empirical validation is the development of a proof of concept, which will be applied to a specific company within a niche industry. The company in focus is "Cascades," a player in the pulp and paper sector.

This proof of concept will not merely remain theoretical; it will be instantiated through tangible algorithmic processes. By leveraging the R programming language, renowned for its statistical and data analysis capabilities, the research will algorithmically orchestrate the exploration process specific to the pulp and paper industry. This hands-on approach ensures that the theoretical constructs discussed are grounded in real-world applicability, offering both academic and practical insights into the synergies between data science and organizational ambidexterity.

The reason behind choosing the pulp and paper industry is to test the extent of the success of our conceptual framework by implementing it on a niche industry where such research methods were never used as far as we know. Hence, providing an even more powerful proof of concept. To do that, we will first collect the twitter data of companies in the pulp and paper industry in addition to our randomly chosen company and shareholder influencers. We will then measure the audience's engagement with their tweets and identify the contents of the tweets that were voted as the most important by the twitter crowd. Thirdly, we will analyse the contents of these tweets and represent our analysis using visuals to come up with insights that will influence decision makers. The aforementioned three-step data science methodology of explorative processes is synonymous to the first three steps in figure 1 of the Introduction chapter. Consequently, we expect our data science approach to reduce the uncertainties embedded in exploration when crafting winning strategies while also minimizing the use of resources. The following flowchart provides a more detailed step-by-step summary of our methodology.

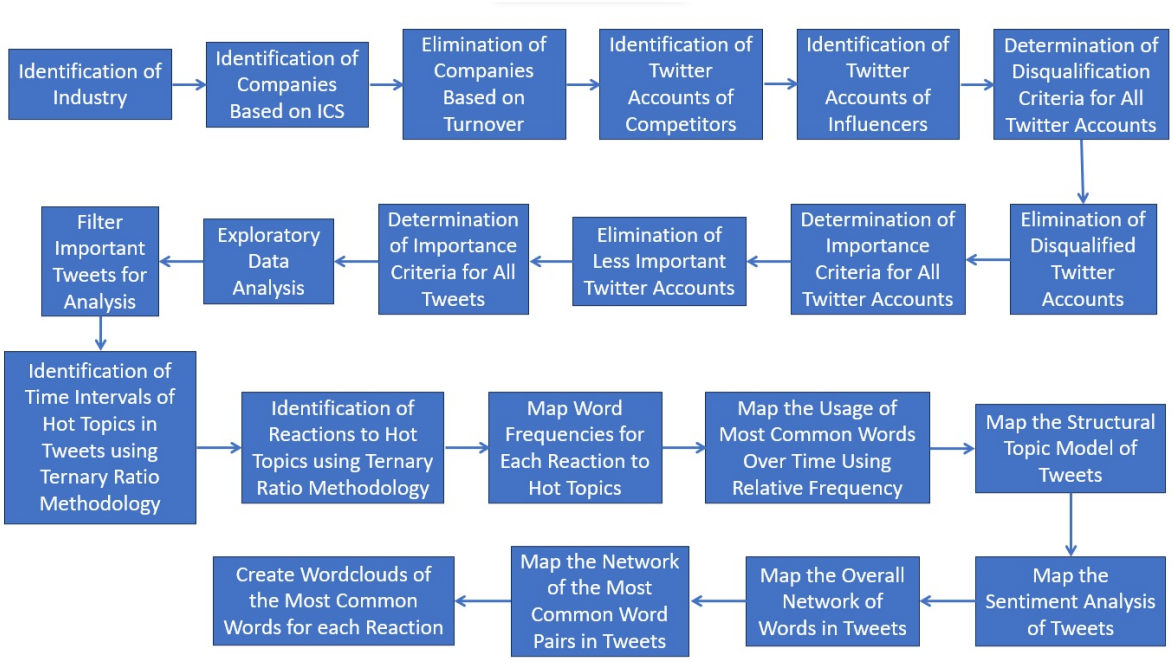


Figure 24: A step-by-step breakdown of our methodology. All of these steps will be taken using coding in R

The overarching objective of this research is to establish a comprehensive, radar-like system designed to capture high-velocity signals emanating from Online Social Networks (OSNs). This system isn't just about data collection; it's about transforming raw data into actionable intelligence. Through sophisticated analysis, these signals will be distilled into meaningful insights, which will then be visually represented in a manner conducive to strategic decision-making. In essence, this system aims to bridge the gap between the vast, dynamic digital chatter of OSNs and the strategic imperatives of organizations, ensuring that decisions are informed, timely, and aligned with the evolving digital landscape.

Our methodology aims to reduce the difficulties accompanying exploration without eliminating its proven advantages, such as the case discussed about the diffusion of process management techniques to exploratory activities leading to dampened innovation (Benner and Tushman 2003). Indeed, our conceptual framework outlines how that could be achieved using data science to transform conventional exploration into efficient exploration in order to eliminate the challenges of ambidexterity and take advantage of its benefits. Starting with the first challenge, data science will allow organisations to collect, filter through and analyse tweets of our three main stakeholders: competitors, shareholders and customers. Considering our intent of reducing that first challenge, our data science approach to exploration centralizes these three processes. The result of this activity centralization is faster decision-making,

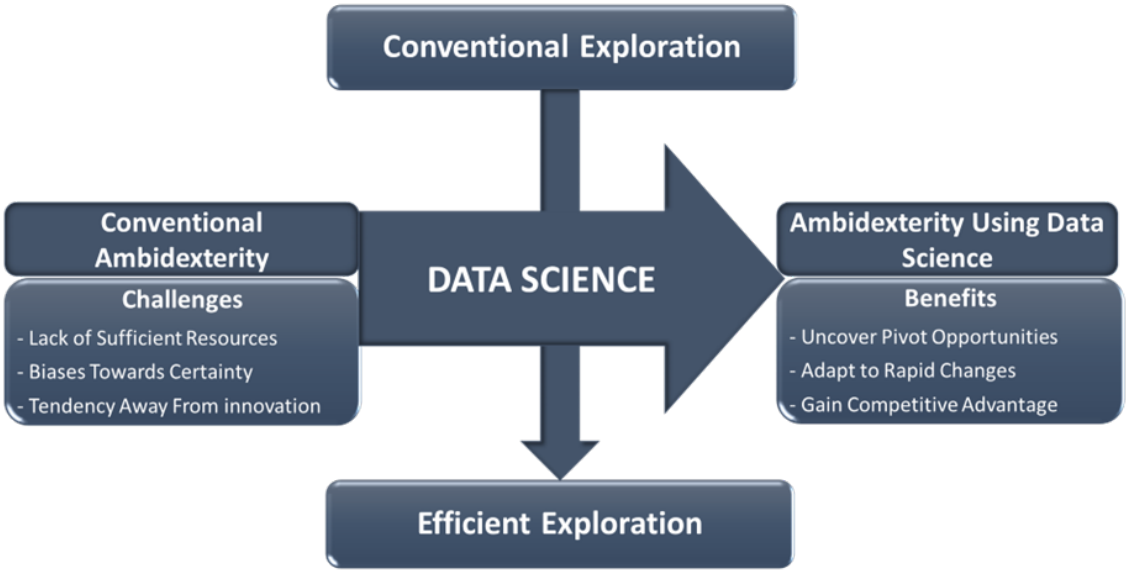


Figure 25: The conceptual framework summarizing the premises of our methodology

better coordination of resources as well as more effective monitoring of any nuances in the incoming signals which eventually lead to a competitive advantage in the operating model.

Hence, the organisations working on becoming ambidextrous using data science could have at least the same amount of productive knowledge as their peers in the industry without employing the time-consuming and expensive conventional methods of exploration discussed in the literature review. Also, our proposed method would ease the biases towards certainty and allow exploitation to occur without sacrificing large amounts of resources for exploration - the core of this dissertation. Hence, minimizing the paradox of administration (Thompson 2003). It is important to note that cleaning up the collected twitter data, better known as data wrangling in the data science world, is more than 80% of the battle and this is where the biggest efforts are directed throughout the methodology chapter.

Moreover, data science would provide a tool for organisations to foresee upcoming changes before they happen in order to create the tide rather than ride it. In terms of innovation, the Natural Language Processing (NLP) techniques along with the network analysis to be employed by organisations seeking ambidexterity using data science, will allow them to build upon innovations commenced by their peers in the industry rather than start from scratch in order to gain the unique competitive advantage they are seeking. Finally, using data science to map out and analyse the industry network of stakeholders will provide more clarity,

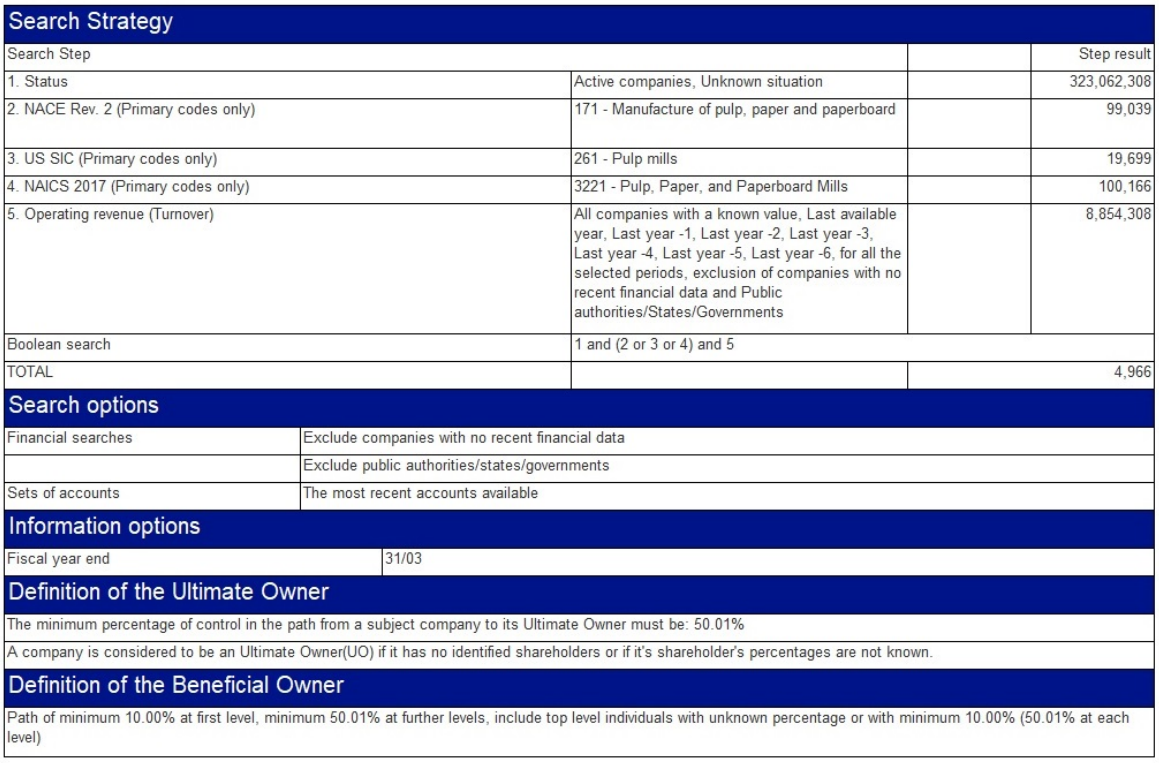
identifying major clusters and stakeholders that could lead towards the competitive advantage we are referring to. Consequently, this would map the competitive landscape and highlight any pivot opportunities which may lead to virtuous cycles in their business models. This research protocol along with the observation data collection technique described, highlight how data science can simplify exploratory activities and provide real competitiveness results.

After evaluating the final list of twitter accounts for analysis, our methodology will be split into 3 complementary parts that allow us to achieve our ultimate goal: the ternary ratio methodology, Natural Language Processing (NLP) techniques, followed by network analysis.

4.1 Introduction to the Pulp & Paper Industry

The pulp and paper industry is made up of manufacturing companies that transform woody or plant material into pulps, paper and paperboards (Kuhlberg 2015). Knowing that, we searched in ORBIS Bureau Van Dijk for all companies in the database that fit this definition. We used NACE primary code 171 which represents manufacturers of pulp, paper and paperboard and the result was 99,039 enterprises. We also factored in US SIC primary code 261 which represents pulp mills and the result was a total of 19,699 companies. Moreover, we included NAICS primary code 3221 which represents pulp, paper and paperboard mills and the result of that was 100,166 companies. Using the industry classification systems was helpful to ensure we are as inclusive as possible in our search and to avoid any biases due to the unintentional omittance of other companies in the pulp and paper industry. We also added the constraint of only having results on companies with a known value of turnover, with the last turnover available in the year 2015 onward, excluding companies with no recent financial data or any public entities. These constraints brought down our total number of pulp and paper companies to 4,966.

Turnover was chosen as a deciding factor of which companies to analyse for three main reasons. First, it is one of the most universally used factors to gauge and compare companies' performance over time (Chizobah 2022). Second, it is a numerical value that clearly contrasts how fierce the competition on market share is between companies when looking at how close their values of turnover are. And third, the pulp and paper industry is very price elastic on the demand side in the sense that companies would lose revenue and market share if they raise their prices (Karikallio, Mäki-Fränti, and Suhonen 2011). This implies that any change in turnover of any company in this industry is most likely due to gaining more customers rather than raising prices. These new customers could be gained either from competitors or



Search Strategy			
Search Step			Step result
1. Status	Active companies, Unknown situation		323,062,308
2. NACE Rev. 2 (Primary codes only)	171 - Manufacture of pulp, paper and paperboard		99,039
3. US SIC (Primary codes only)	261 - Pulp mills		19,699
4. NAICS 2017 (Primary codes only)	3221 - Pulp, Paper, and Paperboard Mills		100,166
5. Operating revenue (Turnover)	All companies with a known value, Last available year, Last year -1, Last year -2, Last year -3, Last year -4, Last year -5, Last year -6, for all the selected periods, exclusion of companies with no recent financial data and Public authorities/States/Governments		8,854,308
Boolean search	1 and (2 or 3 or 4) and 5		
TOTAL			4,966

Search options	
Financial searches	Exclude companies with no recent financial data
	Exclude public authorities/states/governments
Sets of accounts	The most recent accounts available

Information options	
Fiscal year end	31/03

Definition of the Ultimate Owner
The minimum percentage of control in the path from a subject company to its Ultimate Owner must be: 50.01%
A company is considered to be an Ultimate Owner(UO) if it has no identified shareholders or if it's shareholder's percentages are not known.

Definition of the Beneficial Owner
Path of minimum 10.00% at first level, minimum 50.01% at further levels, include top level individuals with unknown percentage or with minimum 10.00% (50.01% at each level)

Figure 26: Search strategy of companies in the pulp and paper industry

from completely new industries, i.e. pivot opportunities. The stakeholders we believe could have an effect on turnover are: customers, shareholders and competitors.

To operationalize the research objectives, the study will employ R programming to delve into the digital conversations of the three previously identified stakeholders: customers, competitors, and shareholders. The primary metric for gauging the resonance of these conversations on Twitter will be the level of engagement, quantified through three distinct activities: replies, retweets, and likes.

For the customer segment, the focus will be twofold. Firstly, we will analyze the engagement metrics for conversations related to our primary company of interest, Cascades. Simultaneously, to ensure a holistic understanding of the market dynamics, we will also monitor the engagement metrics for conversations related to Cascades' competitors. This dual approach ensures that we capture a comprehensive snapshot of the customer sentiment and engagement landscape within the pulp and paper industry.

Furthermore, to provide a more rounded perspective, the study will also measure engagement levels associated with news articles or mentions related to our chosen organizations. This will offer insights into the broader public perception and the potential impact of media narratives

on stakeholder engagement. Through this methodical approach, the research aims to harness the power of R programming to transform vast OSN data into actionable strategic insights for organizations.

As for shareholders, they will be proxied by the twitter accounts of shareholder influencers such as environmental lobby groups or business news sources, specifically the "Bloombergs" of pulp & paper. The tweets of shareholder influencers are ideal to identify upcoming trends and test the validity of the links between shareholders and customers. But most importantly, we will observe Cascades' competitors on twitter to detect insights within their tweets or their tweets' engagements and ensure that Cascades are always up-to-date with their competitors. We are aware that some engagements may be performed by shareholders or other competitors in addition to the public instead of the customers we are targeting, but our assumptions are that either these engagements make up an insignificant portion of the total engagements based on the data we have seen, or much of the crowd that engage with these organisations on twitter actually use their essential products in every household which eventually makes them customers. It is important to highlight that enlarging the listening or detection radius to include twitter accounts other than the 3 stakeholders, that are not directly related to the pulp and paper industry, could be useful for the exploration process. Fujifilm's zoom out from the photography business and their discovery that they are also in the chemical business to create ASTALIFT is a case in point (FUJIFILM 2007).

To summarize our approach, customers in our analysis are represented as the total number of engagements or activities towards a tweet. Shareholders are represented by the twitter accounts of shareholder influencers. And competitors are represented by their own respective twitter accounts.

4.2 Data Collection

Twitter rest API will be our collection method of the tweets of the stakeholders with their meta data within our defined radius. API stands for Application Programming Interface. Twitter API allows access to acquire publicly available data on twitter in a programmatic way (Twitter 2023b). This enables the batch processing of large amounts of Twitter data. The data and metadata obtained includes, but is not limited to: date and time of tweet, twitter identifier, tweet, number of likes, number of replies, number of retweets, geolocation of tweet (if location services were turned on for twitter), URL of tweet, links to photos or videos attached to tweet, hashtags... etc. This data can be accessed and downloaded in csv format

by obtaining access from twitter and specifying few pieces of information such as the twitter identifier being targeted over a certain period of time.

From an ethical standpoint, this data is considered as primary data and its collection falls within the ethical boundaries for two reasons: Firstly, this data was made public by their owners through tweets on the world wide web. And second, the owners of these tweets had already signed the privacy agreement provided by twitter and any use of this data naturally inherits the privacy agreement of twitter that states: "Twitter is public and Tweets are immediately viewable and searchable by anyone around the world" (T. Twitter 2020). This encouraged putting more focus on the quality of the data on which we will build upon our research.

There are two general types of Twitter API: rest API and streaming API. Rest API, which was used in this thesis, presents a snapshot in time of historical twitter data and metadata of the specified account(s) based on specified parameters. Whereas streaming Twitter API allows getting the twitter data and metadata in realtime as they happen (Twitter 2023a). Ideally, streaming Twitter API should be used to influence decision makers since it is based on up-to-date realtime data. But for the sake of this dissertation, we used rest API to provide a proof of concept.

After that, we will use data science to filter out the unproductive knowledge from the obtained tweets based on justified preset conditions before analyzing the remaining data. Finally, we will rearrange our filtered data and create insightful visuals to influence decision makers. These steps will be detailed in the subsections below.

4.3 Identification of Twitter Accounts for Analysis

Several published studies have shown that companies post on social media to exhibit their latest and greatest, to gain legitimacy and maintain a healthy customer service (Kouloukoui et al. 2023; Lundmark, Oh, and Verhaal 2017; Mette Morsing, Schultz, and Nielsen 2008; M. Morsing and Beckmann 2006). We are aware that competitors will not tweet the blueprints of their latest innovation for example, but those reputable cited papers imply that companies will still tweet about their latest updates in order to achieve the three goals we mentioned earlier. Ultimately, we reemphasize the point that our methodology allows us to have the same amount of information as the market has in order to create a pareto optimal kind of balance. The two subsections below explain how we identified twitter accounts of competitors

followed by twitter accounts of shareholder influencers.

4.3.1 Twitter Accounts of Competitors

Since Cascades ranks in 16th place in terms of turnover out of those 4,966 companies, and since it would be very difficult for any company to advance multiple ranks in one shot due to the high price elasticity, we decided to just include 4 companies ranked below Cascades along with all the other companies ranked above in turnover. The difference in turnover between Cascades and the 4 companies ranked below it does not exceed 3% rank-over-rank, until we reach a Taiwanese pulp and paper company called YFY in 21st place which shows a difference in turnover of around 13% with the company ranked above it. Hence, YFY Inc., the 5th company ranked below Cascades, was not included in our analysis since the difference in turnover between YFY and the 4th ranked company below Cascades is around 13% which reduces its likelihood of surpassing it in a short period of time. In other words, it would not be necessary to include YFY Inc. in our detection radius since it is unlikely that they would pose a direct threat on Cascades knowing the industry dynamics.

We then took those companies and searched for their twitter accounts or the twitter accounts of their executive team members. Few inclusions were made without following our preset conditions above, namely: SD Waste and Elopak Pure-Pak. Even though SD Waste does not qualify to be part of these companies based on their turnover, they were still included in the analysis due to their collaboration with several of the companies we included. As for Elopak Pure-Pak, they were included because they entered into a licencing agreement with Nippon Paper who ranks in 7th place in terms of turnover (Cornall 2022). These out-of-scope inclusions were made to ensure we do not miss out on any potentially valuable signals from their collaboration with the top pulp and paper companies we included based on turnover.

It should be noted that some of the companies we included for our analysis are not present on twitter. To remedy that, an extensive google search for their twitter accounts was performed without any success. We also tried searching for their board members or executive teams individually on twitter without any success. After investigating, most of these companies turned out to be from China or Japan and it is well-known that there are other large social media platforms used there such as Line and Ameba that compete with twitter (Harding 2010). In addition, the concept of being present where your audience is and the difference in business and marketing culture may have played a big role in why these companies are not present on twitter, as could be justified by Ghemawat's CAGE framework (Pankaj.

Ghemawat 2007). As for our organisations of interest who are present on twitter, it is important to note that some of them tweet using multiple twitter accounts. These accounts may engage the audience in different languages, or may represent subsidiaries of the same company such as the case of UPM who has a Finnish account, or Indah Kiat which is part of Asia Pulp & Paper. Regardless, we still included all these accounts or organisations for now to be as comprehensive as possible and to avoid any biases before our analysis. As a result, we end up with a total of 22 pulp and paper companies represented by 21 twitter accounts.

4.3.2 Twitter Accounts of Shareholder Influencers

We also went through more than 60 different twitter accounts from the "You might like" recommendation window on twitter located on the right side to identify potential influencer candidates based on Twitter's recommendation algorithm. We terminated our search on the "You might like" list when we exhausted all recommendations made by Twitter's algorithm, i.e. not getting any new twitter account recommendations. Another way of finding influencer accounts was through searching for twitter accounts that mentioned the companies we included in our analysis. This ensures we maintain a clear overview of the network of tweets linking the different stakeholders through replies, retweets and likes. Put differently, the influencers that did not come up in our extensive search do not affect our results in any way since they are not connected directly or indirectly to our chosen pulp and paper companies through a tweet or engagement and thus, it is very unlikely for these excluded influencers to have any effects on shareholders planning to invest in our included companies. As a result, we end up with a total of 16 influencers represented by 16 twitter accounts.

Using an aggregate lens, the twitter accounts of both the competitors and influencers we identified total up to 167,287 tweets covering dates from March 2009 until November 2022. Throughout the following sections, we will eliminate several of these 37 twitter accounts based on justified criteria for our analysis. The table on the next page shows a summary of the pulp and paper companies in addition to the influencers we included for our analysis so far. The data in this table was obtained in December of 2022. It is important to highlight that the few companies sharing the same turnover value in this table ultimately refer to the same entity.

Last Available Turnover (Million USD)	Organisation Name	Twitter Identifier	Language of Tweets	Total Number of Followers	Total Number of Tweets	Organisation Type
19,363	International Paper	@IntlPaperCO	English	9,219	5,919	P&P Company
12,468	Amcor Packaging	@amcorpackaging	English	N/A	N/A	P&P Company
12,011	Oji Holdings	N/A	N/A	N/A	N/A	P&P Company
11,819	Stora Enso	@storaenso	English	12,000	2,372	P&P Company
11,268	UPM	@UPMGlobal	English	11,900	7,044	P&P Company
11,268	UPM Suomi	@UPMSuomi	Finnish	4,532	2,147	P&P Company
8,747	Mondi Group	@mondigroup	English	3,122	630	P&P Company
8,538	Elopak Pure-Pak	@Pure_Pak	English	2,662	1,129	P&P Company
8,538	Nippon Paper	N/A	N/A	N/A	N/A	P&P Company
7,638	Suzano	@Suzano_	Portuguese	N/A	N/A	P&P Company
6,626	APP Sinar Mas	@AsiaPulpPaper	English	4,882	6,874	P&P Company
6,626	Asia Pulp & Paper ID	@AsiaPulpPaperID	Indonesian	1,751	4,848	P&P Company
6,470	Somos CMPC	@SomosCMPC	Spanish	3,065	792	P&P Company
5,308	Shanying	N/A	N/A	N/A	N/A	P&P Company
5,265	Sappi Europe	@SappiEurope	English	3,875	1,366	P&P Company
5,265	Sappi Group	@SappiGroup	English	1,991	699	P&P Company
5,265	Sappi North America	@SappiNA	English	4,663	3,185	P&P Company
5,265	Sappi Southern Africa	@SappiSouthernA	English	1,058	842	P&P Company
5,003	Daio Paper	N/A	N/A	N/A	N/A	P&P Company
4,272	Lee & Man Paper	N/A	N/A	N/A	N/A	P&P Company
3,783	SCG Packaging	@SCGP_official	Thai	N/A	N/A	P&P Company
3,770	Cascades	@CascadesSD	English	2,258	2,392	P&P Company
3,668	Domtar	@DomtarEveryday	English	4,790	10,000	P&P Company
3,668	Domtar Corporation	@DomtarCorp	English	1,959	322	P&P Company
3,664	Resolute FP	@resolutefp	English	5,916	3,484	P&P Company
3,562	MM Karton	N/A	N/A	N/A	N/A	P&P Company
3,546	Indah Kiat Pulp & Paper	N/A	N/A	N/A	N/A	P&P Company
N/A	SD Waste Paper Recycling Centre	@SdWaste	English	477	255	P&P Company
N/A	AcuComm	@AcuComm	English	1,392	3,888	Influencer
N/A	Business Wire-NR	@BW_NaturalResou	English	N/A	N/A	Influencer
N/A	Fastmarkets Forest Products	@FastmarketsFP	English	7,518	36,200	Influencer
N/A	Inside Packaging & Packaging Gateway	@InsidePackaging	English	16,300	5,113	Influencer
N/A	Mighty Earth ðŸŒ	@StandMighty	English	9,186	4,426	Influencer
N/A	Packaging World	@packagingworld	English	36,400	5,501	Influencer
N/A	PackagingNews	@PackNews	English	19,200	15,500	Influencer
N/A	Paper Mart	@Papermart_	English	1,033	1,924	Influencer
N/A	Paper Technology International	@PaperTechnology	English	1,610	745	Influencer
N/A	PAPEREX	@Paperex_Hyve	English	1,220	760	Influencer
N/A	Papnews	@paper_industry_	English	1,412	6,778	Influencer
N/A	PPI Europe	@ppi_europe	English	4,281	3,508	Influencer
N/A	Pulp Paper Energy News	@Vesa_Pulp_Paper	English	71,500	1,169	Influencer
N/A	Smithers	@WeAreSmithers	English	325	349	Influencer
N/A	TAPPI	@TAPPITWEETS	English	2,685	3,911	Influencer
N/A	Woodbizforum	@woodbizforum	English	509	11,100	Influencer

Figure 27: Table showing all the different stakeholders for our analysis

4.4 Elimination Criteria for Twitter Accounts

Taking the 38 organisations we identified that encompass pulp and paper companies along with influencers, we filtered out organisations with private accounts whose tweets cannot be obtained or viewed by non-followers. We also excluded organisations that do not have twitter accounts and whose executives have no twitter accounts either. Furthermore, we eliminated organisations with zero tweets on their twitter accounts. We also excluded inactive accounts with no tweets for at least 3 years on their current or previous accounts since COVID19 hit the economy. The reason behind that last exclusion is we wanted to level the analysis playing field with organisations that were still active on twitter during the pandemic. We then filtered out twitter accounts whose tweets are private. We also filtered out organisations that tweet in any language other than English. Organisations that have multiple twitter accounts tweeting in multiple languages have not been excluded, but their non-English twitter accounts have been excluded in order to qualify for our Natural Language Processing (NLP) analysis. This leaves us with a total of 15 twitter accounts representing 11 pulp and paper companies, and a total of 15 twitter accounts representing 15 influencers. As a result of our eliminations, the total number of tweets for our analysis went down from 167,287 to 57,925 tweets. The table on the next page shows the remaining organisations and their accounts.

Last Available Turnover (Million USD)	Organisation Name	Organisation Type	Twitter Identifier	Language of Tweets	Total Number of Followers	Total Number of Tweets
19,363	International Paper	P&P Company	@IntlPaperCO	English	9,219	5,919
11,819	Stora Enso	P&P Company	@storaenso	English	12,000	2,372
11,268	UPM	P&P Company	@UPMGlobal	English	11,900	7,044
8,747	Mondi Group	P&P Company	@mondigroup	English	3,122	630
8,538	Elopak Pure-Pak	P&P Company	@Pure_Pak	English	2,662	1,129
6,626	Asia Pulp & Paper Sinar Mas	P&P Company	@AsiaPulpPaper	English	4,882	6,874
5,265	Sappi	P&P Company	@SappiEurope	English	3,875	1,366
		P&P Company	@SappiGroup	English	1,991	699
		P&P Company	@SappiNA	English	4,663	3,185
		P&P Company	@SappiSouthernA	English	1,058	842
3,770	Cascades	P&P Company	@CascadesSD	English	2,258	2,392
3,668	Domtar Corporation	P&P Company	@DomtarCorp	English	1,961	322
		P&P Company	@DomtarEveryday	English	4,787	10,000
3,664	Resolute FP	P&P Company	@resolutefp	English	5,916	3,484
N/A	SD Waste Paper Recycling Centre	P&P Company	@SdWaste	English	477	255
N/A	AcuComm	Influencer	@AcuComm	English	1,392	3,888
N/A	Fastmarkets Forest Products	Influencer	@FastmarketsFP	English	7,518	36,200
N/A	Inside Packaging & Packaging Gateway	Influencer	@InsidePackaging	English	16,300	5,113
N/A	Packaging World	Influencer	@packagingworld	English	36,400	5,501
N/A	PackagingNews	Influencer	@PackNews	English	19,200	15,500
N/A	Papnews	Influencer	@paper_industry_	English	1,412	6,778
N/A	PAPEREX	Influencer	@Paperex_Hyve	English	1,220	760
N/A	Paper Mart	Influencer	@Papermart_	English	1,033	1,924
N/A	Paper Technology International	Influencer	@PaperTechnology	English	1,610	745
N/A	PPI Europe	Influencer	@ppi_europe	English	4,281	3,508
N/A	Mighty Earth	Influencer	@StandMighty	English	9,186	4,426
N/A	TAPPI	Influencer	@TAPPITWEETS	English	2,685	3,911
N/A	Pulp Paper Energy News	Influencer	@Vesa_Pulp_Paper	English	1,712	71,600
N/A	Smithers	Influencer	@WeAreSmithers	English	325	349
N/A	Woodbizforum	Influencer	@woodbizforum	English	509	11,100

Figure 28: Table showing the filtered results of the the different stakeholders for our analsysis

4.5 Elimination Criteria for Tweets

In this section, we are going to examine the quality of our data in order to prepare for the statistical analysis of the counts of likes, replies, retweets and their ternary ratios. Let us recall that we ended up with a total 57,925 tweets after our most recent elimination in the previous section of twitter accounts for our analysis. At this point, we will programmatically add 4 new columns in the data and meta data spreadsheet, better known as data frame in the data science world, that we have already obtained using Twitter's rest API. The columns to be added are the total activity for each tweet, the ternary ratio with respect to replies, the ternary ratio with respect to retweets and the ternary ratio with respect to likes. The following sections justify the significance of these additions.

4.5.1 Total Activity for Each Tweet

The first column to be added is the sum of all activity counts of each tweet, referred to as Total Activity. The following formula represents the total activity value.

Total Activity = Total Replies + Total Retweets + Total Likes

The total activity is an important part of our analysis for four reasons: Firstly, it will help us to radically bring down the total number of tweets for our analysis by eliminating unimportant tweets. For example, some tweets might have not garnered any attention at all by our stakeholders. This translates into 0 replies, 0 retweets and 0 likes which adds up to a total activity of 0. This means the twitter crowd decided that those tweets are not important, and thus it would be a waste of resources to include them in our analysis since the topics discussed in those tweets will not bring in any value added to our company of interest, Cascades. That is to say, Cascades will not gain any benefit by analyzing any tweets that are not perceived as interesting by the twitter crowd to ultimately use for their competitive advantage. We are aware that this may be perceived as survivorship bias, but our assumption is that the audience are the real controllers of market dynamics by expressing what they want or need on twitter, synonymous to how supply and demand works from an economic point of view. Hence, adding these "unimportant" tweets will mask or dilute the significance of tweets that were voted as more important by the twitter crowd.

The second reason behind the significance of total activity is to help us create a criterion by setting a threshold to the amount of total activity that makes a tweet important. Put differently, tweets having a total activity of more than 0 but less than a certain threshold are considered as unimportant because if they were actually perceived as important, the twitter crowd would have simply reacted more with those tweets increasing their activity total above the threshold. It is important to note that the answer to what the threshold should be is quite subjective and may depend on the twitter account, their number of followers, their number of tweets, the frequency of their tweets, the way they convey the message, the demographics of their followers... etc. For us, we will explain in detail "how many" makes a tweet important in addition to our rationale in the following section.

The third reason why the value of total activity is important is because it will be used in tandem with the value of ternary ratios of the activity types. Take the following example. Imagine two randomly chosen tweets from our dataframe of 57,925 tweets where both of them have a ternary ratio value of replies = 0.92. At first glance, this would mislead us into thinking that both of those tweets created a lot of conversation, potentially controversial, on twitter. But if we look closely at the activity total of both of these tweets, we will realize that one of them has a total activity of 250 while the other has a total activity of 25. This clearly

contrasts which tweet requires more attention for analysis and protects us from falling into biases traps.

The fourth reason why total activity matters is because it gives us the full picture of a tweet's importance and prevents us from accidentally eliminating any tweets with two activities having a total count of zero while the third activity has a substantial total count.

4.6 Elimination of Less Important Twitter Accounts

Before analyzing our data, we will derive, justify and use the criteria we defined in the previous sections to define what makes a twitter account important for our analysis. Recall that we ended up with 30 twitter accounts for analysis where 15 of them are for pulp and paper companies and the remaining 15 are for influencers. These twitter accounts represent our key stakeholders that we referred to as the major organic players in the pulp and paper industry and who we perceive as having an effect on the market.

twitter name	maximum replies	maximum retweets	maximum likes	maximum activity
paper_industry_	1	1	2	3
papertechnology	1	1	4	4
acucomm	1	5	7	10
fastmarketsfp	1	5	9	13
sdwaste	1	2	12	14
sappieurope	5	7	11	14
woodbizforum	12	7	11	15
cascadessd	3	7	13	17
papermart_	2	7	14	18
paperex_hyve	2	8	11	20
ppi_europe	3	11	15	26
sappina	3	13	18	31
domtarcorp domtareveryday	7	12	25	31
vesa_pulp_paper	4	7	36	43
packnews	5	15	29	48
wearesmithers	3	5	45	50
pure_pak	2	14	38	54
packagingworld	5	18	33	55
asiapulppaper	6	58	25	62
insidepackaging	4	31	32	65
mondigroup	66	8	83	89
sappisoutherna	3	19	78	95
tappitweets	152	6	11	158
resolutefp	32	76	93	176
sappigroup	7	27	215	235
intlpaperco	222	33	86	266
storaenso	5	39	233	277
standmighty	26	157	279	443
upmglobal	47	191	870	1064

Figure 29: Table showing the maximum activity count for each of the 30 twitter accounts we identified as key in the previous section

Taking these accounts and building up on the previous sections, we identified the maximum total activity attained for any tweet made by each of those 30 twitter accounts. The reason behind evaluating the maximum total activity is because it would help us identify which of those 30 twitter accounts garner relatively more engagement to start a conversation and potentially have an effect on the market through their tweets. For example, the shareholder influencer account called `@paper_industry_` has a maximum total activity count of 3 even though they have 6,778 tweets and 1,412 followers. We could compare that to `@wearesmithers` who only have 325 followers and 349 tweets, but whose maximum total activity count reached 50. This negates the assumption we stated earlier that the amount of total activity to tweets may depend on the total number of tweets and followers. In other words, `@wearesmithers` garners more engagement from the twitter crowd than `@paper_industry_` despite having much less followers and total tweets. Based on our research protocol and the twitter crowd preferences, this makes `@wearesmithers` a stronger influencer in the pulp and paper industry than `@paper_industry_`. But to really confirm that, we would need to analyse the nature of the followers of both twitter accounts to see if the engagement garnered by `@wearesmithers` are due to their ego network as compared to `@paper_industry_` who may have a more dispersed and decentralized network of followers. Regardless, we will not venture into finding out as this extends beyond the scope of our proof of concept. We could also use the same generalization on the twitter accounts of the 15 pulp and paper companies we identified as key.

Summing up, we adopted the constraint of maximum total activity count because it is the clearest and least biased criterion in our point of view since it is based on the preferences of the twitter crowd who decide which twitter accounts attract engagements and may in turn have an effect on the market.

Moving ahead in our research approach, we will consistently utilize a selective skimming mechanism to extract the most pertinent and valuable data, ensuring we focus on the most influential and impactful elements. In the context of our current analysis, we've defined our benchmark as the 75th percentile of the highest total activity value across the 30 Twitter accounts detailed in the table. This methodological choice ensures that our analysis remains concentrated on the most significant data points, allowing for a more refined and insightful exploration of the subject matter.

To justify our preference from a statistical point of view, the total activity values are classified

as discrete since they represent sum of counts. In addition, our individual activity data points are all positive and their distribution is not normal, showing heavy right skewness for any of the three activity counts we are looking at including the total activity. At the same time, there are lots of total activity outliers that should not be eliminated because these tweets particularly matter to our analysis. One example of this is an influencer account called `@tappitweets`, having a mean replies value of 0.25 for all their tweets and a maximum replies value of 152. Eliminating these outliers or tampering in anyway with the distribution of their activity count data will eliminate important information that we are looking for and will prevent us from reaching our objective of finding the tweets and topics voted as the most important by the audience. That being said, we will use median and 3rd quartile to identify any hot topics instead of using mean and standard deviation which do not fit the nature of our data.

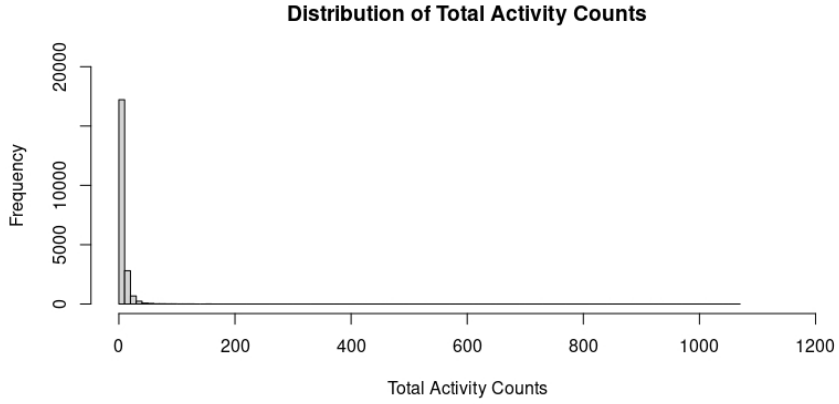


Figure 30: Histogram showing the non-normal distribution of total activity counts of the different tweets

We are aware that choosing 75^{th} percentile is subjective and could eliminate twitter accounts who have a slightly less maximum total activity than the threshold we set. But our goal in this dissertation is to provide a proof of concept, while in reality, an algorithmic trial and error system could be programmed to find the optimum results for an AI-chosen threshold. On the other hand, we are also aware that enlarging our detection radius could include less important twitter accounts that may create extra noise in our analysis even though they have a maximum total activity higher than the 75^{th} percentile. Overall, this could be justified also because we can see that the first major step up in % change found after the median in the maximum total activity happens at the 3rd quartile due to how our data is distributed in figure 29.

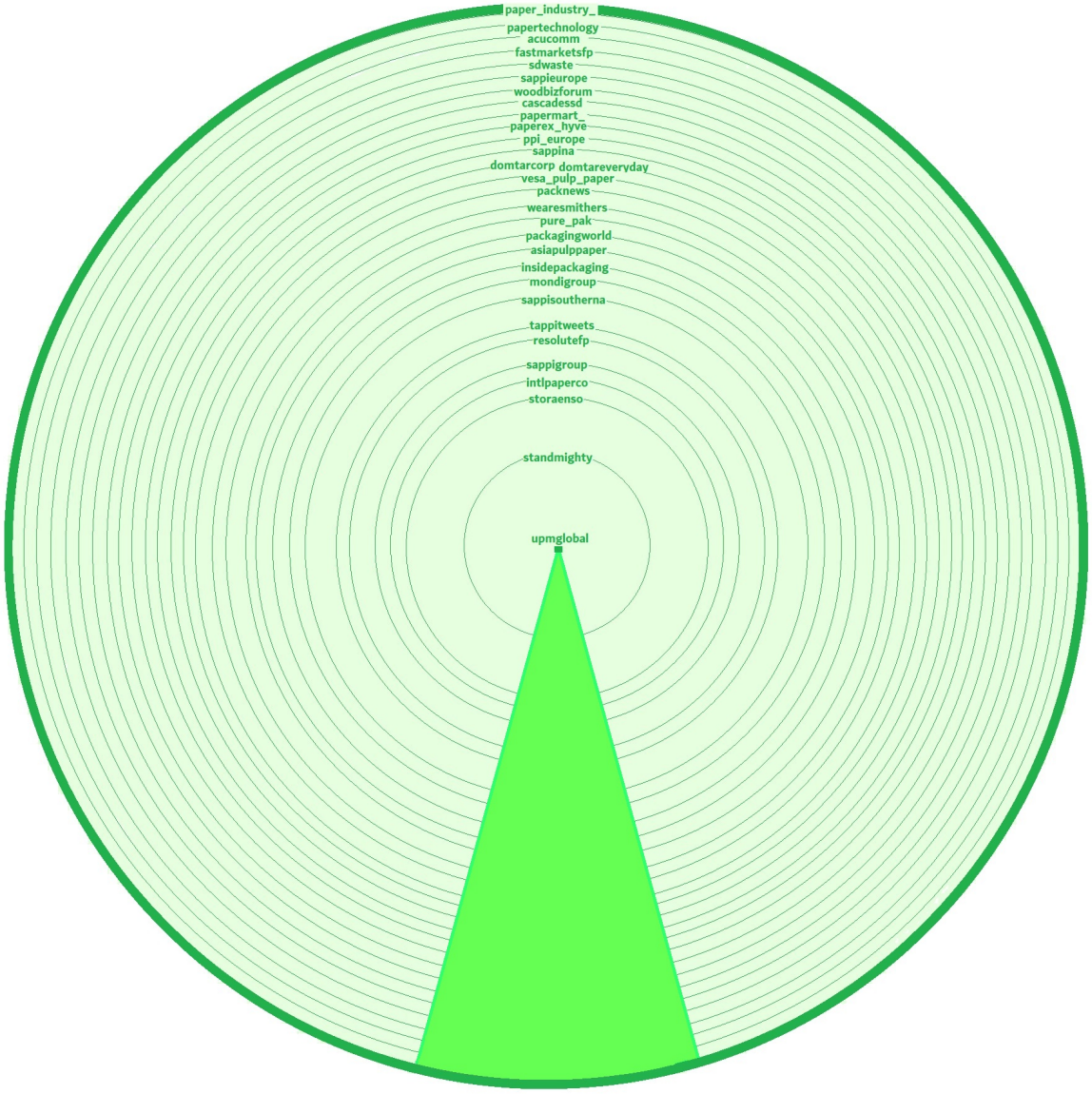


Figure 31: A to-scale radar figure showing the importance of twitter accounts based on maximum total activity count. The importance of twitter accounts increases as the radius decreases. All twitter accounts shown within radar circle and placed on different detection radii represent key stakeholders in the pulp and paper industry. All other twitter accounts outside our radar may represent pivot opportunities. As the maximum total activity attained by a twitter account decreases, the detection radius to include this twitter account increases

Figure 31 on the previous page depicts how the maximum total activity count, which is our twitter account importance threshold, acts as the range of our radar detection radius. And just like in any defense or attack system, the priority is to focus on the objects that appear to be closer to the center. As we move further away from the center, the maximum total activity count decreases which implies that the importance of the twitter account decreases due to its low perceived effect to engage the twitter crowd. Hence, its lower effect on market dynamics. After all, our goal is to detect and analyze the important high speed signals from OSNs in order for Cascades to use in a strategic way leading to a competitive advantage.

In our case, the value of the 75^{th} percentile of the maximum total activity is 93.5 , which is slightly below that of `@sappisoutherna` as shown in figure 29. Therefore if we eliminate all the twitter accounts whose maximum total activity attained is less than 93.5, we are left with 9 twitter accounts representing 8 organisations of which 7 are pulp and paper companies and 2 are influencers. Those twitter accounts that garnered the most amount of attention are: Sappi Southern Africa (58% of total tweets are available for analysis), TAPPI (78% of total tweets are available for analysis), Resolute Forest Products (70% of total tweets are available for analysis), Sappi Group (59% of total tweets are available for analysis), International Paper (65% of total tweets are available for analysis), Stora Enso (83% of total tweets are available for analysis), Mighty Earth (61% of total tweets are available for analysis), UPM Global (87% of total tweets are available for analysis) and we will include our company of interest Cascades (91% of total tweets are available for analysis).

There are two reasons behind not being able to obtain 100% of the total tweets for our analysis: Firstly, some accounts have employed the twitter circle where only selected accounts can see specific tweets instead of them being public (Twitter 2022). Secondly, Twitter imposes a cap on how many tweets can be obtained over a certain period of time using the rest API (Twitter 2023b).

This brings our total number of tweets for analysis from 57,925 to 21,189 tweets.

4.7 Elimination of Less Important Tweets

Before separating the remaining 21,189 tweets by individual twitter accounts for our analysis, let us recall our skimming strategy using the following example: if one tweet has an activity total = 1 and another tweet has activity total = 176, both of them may have a ternary ratio = 1 depending on the type of engagement they garnered. But in reality, the tweet with

Rank in Turnover	Last Available Turnover (Million USD)	Organisation Name	Organisation Type	Twitter Identifier	Total Number of Followers	Total Number of Tweets	% of Total Number of Tweets for Analysis
1	19,363	International Paper	P&P Company	@IntlPaperCO	9,219	5,919	65%
4	11,819	Stora Enso	P&P Company	@storaenso	12,000	2,372	83%
5	11,268	UPM	P&P Company	@UPMGlobal	11,900	7,044	87%
11	5,265	Sappi	P&P Company	@SappiGroup	1,991	699	59%
				@SappiSouthernA	1,058	842	51%
16	3,770	Cascades	P&P Company	@CascadesSD	2,258	2,392	91%
17	3,664	Resolute FP	P&P Company	@resolutefp	5,916	3,484	70%
N/A	N/A	Mighty Earth	Influencer	@StandMighty	9,186	4,426	61%
N/A	N/A	TAPPI	influencer	@TAPPITWEETS	2,685	3,911	78%

Figure 32: A summary of the qualifying 75^{th} percentile twitter accounts that will be prepared for analysis

activity total = 176 is much more important for our analysis than the tweet with activity total = 1 due to the amount of attention it attracted from the twitter crowd. That being said, we should be eliminating the tweet with activity total = 1 and ternary ratio = 1 but keep the tweet with activity total = 176 and ternary ratio = 1. So how do we decide what is the threshold of activity total for each twitter account that contrasts whether a tweet is important or not? It really goes by the saying, if everything is important then nothing is important. Again, the threshold is subjective and could change depending on how we define importance.

In our point of view, which we believe to be the best fit and most comprehensive, we pick the 75^{th} percentile of activity total because we want to look at conversations that were voted as most important by the crowd, knowing that the data points are distinct. Taking this threshold is not biased because the twitter crowd took the decision. In other words, if the disqualified tweets having activity total less than the 75^{th} percentile were deemed important by the crowd, they would have garnered more attention which gets translated into more activity putting them potentially in that 75^{th} percentile. This helps us eliminate the misleading noise with our sole dependence on ternary ratio spikes.

4.8 The Ternary Ratio Methodology

We will not just depend on the individual or total activity counts per tweet, we will also take into account the type of activity and what it entails. In other words, a tweet with a relatively high number of replies is implied to be a controversial topic or is seen to require more than just a simple reaction such as a like (Minot 2022). On the other hand, a retweet is implied as an endorsement of the tweet or its replies (Minot 2022). This granularity in the analysis gives

us a clearer idea on the nature of the engagement, just like watching the body language or the face expressions of people communicating to identify how they feel besides listening to what they say. And for that purpose, we will employ the ternary ratio methodology. In other words, the ternary ratio method allows us to understand what our competitors do well, which would garner likes and retweets. This method also identifies what our competitors supposedly do not do well, which would garner controversy on the internet in the form of replies from the twitter audience.

$$R_T(t) = \frac{N_T(t)}{N_{retweets}(t) + N_{likes}(t) + N_{replies}(t)}$$

As can bee seen from the formula above, the ternary ratio provides a normalized volume of user activities in response to a tweet at a specific period in time. Said differently, the numerator would represent the count of one type of activity to a tweet, i.e. replies, retweets or likes, divided by the total count of all 3 types of activities to the same tweet.

4.8.1 Ternary Ratios for Replies

In addition to the total activity column we added to the data frame, we will add be the value of the ternary ratio of the tweets with respect to replies. N(t) represents the count of an activity at a particular moment in time and R(t) in the equation below represents the ternary ratio value of this tweet in relation to replies at a snapshot in time, which is the time we downloaded the tweet metadata.

$$R_{Treplies}(t) = \frac{N_{replies}(t)}{N_{retweets}(t) + N_{likes}(t) + N_{replies}(t)}$$

4.8.2 Ternary Ratios for Retweets

The third column we will add is the value of the ternary ratio of the tweets with respect to retweets. N(t) represents the count of an activity at a particular moment in time and R(t) in the equation below represents the ternary ratio value of this tweet in relation to retweets at a snapshot in time, which is the time we downloaded the tweet metadata.

$$R_{Tretweets}(t) = \frac{N_{retweets}(t)}{N_{retweets}(t) + N_{likes}(t) + N_{replies}(t)}$$

4.8.3 Ternary Ratios for Likes

Finally, the fourth column we are adding represents the value of the ternary ratio of the tweets with respect to likes. N(t) represents the count of an activity at a particular moment in time and R(t) in the equation below represents the ternary ratio value of this tweet in relation to likes at a snapshot in time, which is the time we downloaded the tweet metadata.

$$R_{Tlikes}(t) = \frac{N_{likes}(t)}{N_{retweets}(t) + N_{likes}(t) + N_{replies}(t)}$$

4.9 Natural Language Processing Techniques

Natural Language Processing (NLP) is a subset of data science, as illustrated in the Venn diagram of figure 4 presented in the introduction. NLP, at its core, leverages statistical methods, much like various other artificial intelligence algorithms. This computational approach enables machines to interpret, understand, and generate human language in a manner that is both meaningful and contextually relevant.

This technique algorithmically enables computers to make sense of words, phrases or even large compilations of documents at a level almost as good as humans (IBM 2020). Naturally, NLP is still in its infancy but is improving rapidly with different data scientists and researchers working on it from different places around the world. Few common examples of NLP that are used every day by many humans include Siri on iPhones through speech-to-text algorithms, WhatsApp's or Google's automatic complete predictions when we type a message or use the search engine, and Meta's marketing ads based on sentiment analysis. In our proof of concept, we will utilize two types of NLP on the final list of tweets from the 9 twitter accounts that qualified for our analysis, namely: Structural Topic Modelling and Sentiment Analysis.

It is important to note that we are applying this technique to analyse unstructured data in the form of tweets. This means that the data is unlabeled and has a very flexible structure among other characteristics. The following example is a simplified explanation of what being unlabeled and having a flexible structure means. If we consider the following phrases: Lara goes to school, an apple was eaten quickly by Rula, an argument broke out between Haneen and Amelia, Flora is pretty. The fact that this unstructured data, in the form of phrases, is unlabeled means that each word in those phrases is not classified as a verb, noun, adverb, adjective, subject, object... etc. But we as humans can still understand which word represents

the doer of the verb, the names of the humans presented in the phrases and all the other components of the phrases.

Coming to the flexibility of the structure, the subject or doer of the verb can be at the beginning of the phrase, at the end of the phrase or anywhere in between. So can the object of the sentence. This means that a simple English phrase does not have a fixed structure that starts with a subject followed by a verb then an object. But we as humans can still make sense of the sentence regardless of how it was written or what structure it took.

With all that said, NLP helps machines learn and understand the labels and the structure of the unstructured data so that it could automatically classify them correctly with more accuracy when we present a new phrase. Thus, giving us an understanding of the message being conveyed. It is important to highlight that NLP is a core yet small part of data science and the examples we provided are simplified versions of the techniques that extend way beyond the scope of our proof of concept.

4.9.1 Structural Topic Modelling

Structural topic modelling is a branch of NLP that takes unstructured data as input and produces a list of words, known as tokens in the data science world, as output that statistically appear to be closely related to each other to form a topic (Kouloukoui et al. 2023). Latent Dirichlet Allocation, or LDA, is a very popular way to perform structural topic modelling. LDA has two basic rules: Firstly, it treats different documents as a mixture of topics and secondly, the different topics within these documents are treated as a mixture of words. What LDA does is it outputs a predefined number of words grouped into a predefined number of categories based on statistical probabilistic analyses of their occurrence (Blei 2003). Simply stated, the output of structural topic modelling would be a bar chart showing the probability of the appearance of each word in each topic, depending on the number of words and topics predefined by the user. A common example to explain LDA is if we use it to analyse multiple news articles published by a news agency, where we define the output to have the 10 most common words within just two topics or categories. The output produced would include words such as percent, million, billion, company and more words in category 1. On the other hand, category 2 would include words such as president, government, republican, elections... etc. These output examples tell us that this agency publishes news articles that discuss finances as shown in category 1 and politics as shown in category 2. Specifying the optimum number of categories and words in the output would give a clearer and more

encompassing picture of the messages that the unstructured data is presenting. The advantage of LDA is that it facilitates the batch processing of large numbers of unstructured data and it allows overlapping of words between multiple topics.

Following the two rules of LDA in our case, the documents would be the tweets represented by the different tweet ids whereas the topics would be evaluated based on the different combinations of the words from our tweets. In our proof of concept for each twitter account, we specified the number of most common words as 10 and we experimented with 2, 3 or 4 topics to identify which one gives a clearer representation of the tweets made by the twitter accounts of interest.

4.9.2 Sentiment Analysis

Sentiment analysis is another branch of NLP that takes a phrase, better known as corpus, and breaks it down into its tokens before outputting whether the said phrase has a positive or negative meaning based on the words that make it up. In simple terms, sentiment analysis has 3 different baskets of words known as lexicons. Each word is labelled as positive or negative in one of the lexicons which is called the `bing` lexicon (Qiu et al. 2011; Zhang, Wang, and Liu 2018). On the other hand, the `nrc` lexicon contains words categorized into different feelings in addition to `bing`'s categorizations such as anger, anticipation, disgust, fear, joy, sadness, surprise and trust (Mohammad and Turney 2011). Whereas the `AFINN` lexicon has words rated from +5 to represent a very positive word, going down to words rated as -5 to represent a very negative word (Nielsen 2011). Depending on the lexicon used throughout, the sentiment analysis algorithm will compare the tokens in the corpus we are analyzing to the built-in words in the lexicon. Then, the different scores of the tokens are aggregated to produce an overall score for the whole corpus.

While sentiment analysis may not be always insightful on the tweets made by the 7 twitter accounts of pulp & paper companies or the 2 shareholder influencer accounts we identified, it would be more useful to apply it to the replies and retweet texts produced by the engagements of the twitter crowd. It is important to highlight that sentiment analysis fails to understand sarcasm in texts produced and is always under improvement by data scientists and researchers. Nonetheless, we included it in our dissertation to provide a proof of concept.

4.10 Network Analysis of Words - A Comparative Methodology

Our aim in network analysis is to take the different tokens produced from our NLP techniques to produce a network visualization of these words in order to recreate the ideas presented by the tweeting accounts. Besides the tokens, network analysis could be strictly performed on the hashtags or twitter identifiers that appeared within the tweets. Our assumption is that network analysis will be very useful to bridge any structural gaps in communication and layout the thoughts in order to compare what Cascades and their competitors are talking about, compare what pulp & paper companies VS. influencers are talking about, evaluate which ideas represented by tokens are central to deliver certain messages and identify clusters of words to examine the relationships between them.

In the following chapter, we will create separate dataframes containing all the qualifying tweets of each of the 9 twitter accounts before performing our analysis and presenting our results of each twitter account for discussion.

5 Results

Following the detailed description and justification of our methodology, this chapter provides an overview of the results and comments for each of the 9 twitter accounts separately. Recall that our goal is to generate a proof of concept for a radar-like system that detects and analyzes high speed signals from OSNs in order for Cascades to identify insights and subsequently use them in a strategic way. Repeating the same analysis for each of the 9 twitter accounts allows us to provide various outcome-dependent notes and comments on how different results could influence decision makers differently. Our notes and comments are neither comprehensive nor exhaustive, but they provide a very good idea on how data science can greatly facilitate the exploration process.

The following provides a quick summary of our methodology in relation to the outline of this chapter for each twitter account: We will first present the results of our exploratory data analysis (EDA) and use them to eliminate the relatively less important tweets. Once done, we will have the tweets that were voted as important by the twitter audience which will be used to create four graphs showing the total activity count in addition to the three ternary ratios over time. These graphs will be practical to isolate the tweets within specific time intervals where we have spikes in the total activity count. We will then relate the spikes shown in the total activity count graph to the ternary ratio graphs representing the different reactions, also referred to as engagements. Connecting these four graphs together will help us understand what stimulated each type of engagement.

Recall that the time intervals where tweets show spikes actually translate into hot topics as decided by the twitter crowd through their different engagements in the form of replies, retweets or likes. It is important to note the difficulty in completely isolating the pure stimulants of a certain engagement since it is uncommon to see tweets having only one type of reaction. For that reason, we will take the extra step and use the ternary ratio graphs to specify the type(s) of reaction for each spike in total activity to better understand the underlying reasons behind the reactions. Eventually, we will perform and produce separate results for the three different reactions: replies, retweets and likes.

After that, we will clean up these tweets by excluding entities that have no effects on the overall meaning such as different forms of retweet signalization (e.g: RT), other twitter identifiers preceded by @, punctuation (e.g: ?, !... etc), numbers (e.g: 2021, 350... etc), html links (e.g: https://www....), non-alphanumeric characters (e.g: ¾, †... etc), special

characters (e.g: £, ™... etc), accented letters (e.g: Ã, â... etc) and a list of 1,165 stop words (e.g: and, but, or... etc). It is important to highlight that the presence of these excluded entities will affect our statistical analysis which may in turn mask or dilute the significance of the more important terms we are looking for.

Once the tweets are clean and ready for analysis, we will commence text mining techniques by breaking down the clean phrases of the "spiky" tweets into words which are referred to as tokens in the data science world. What follows is an evaluation of the frequency of the most common words used within those clean tweets in an attempt to classify which words stimulated what reactions. We will then present these stimulating words in the form of three separate frequency bar charts for each type of twitter engagement. We will also create a graph showing the evolution of the use of the most common words over time in order to add some color to the context of their use and make a comparison between the evolution of interests of the different twitter accounts. For that last graph, we will use relative frequency in order to underscore how important the most common words are in relation to the other words used within the same time intervals.

Subsequently, we will employ the NLP techniques we described in the methodology chapter before introducing our network analysis. It is important to highlight that we will integrate in our sentiment analysis different forms of negating words such as not, without, never, no, can't, don't and won't to avoid any potential misinterpretation of the intended meaning (e.g: Avoid interpreting "I am not happy" as a positive sentiment since our algorithm may consider the word "happy" without considering "not"). Finally, we will present a wordcloud grid of the most stimulating common words for each of the three reactions to the different tweets of the 9 twitter account in order to compare the different topics that were tweeted about. As mentioned at the beginning, the process outlined here will be repeated for each one of the 9 twitter accounts.

It is important to highlight that since our company of interest is Cascades, the presented analysis in their section will naturally focus on two out of three stakeholders only: customers and shareholders. Whereas the results of the other twitter accounts will be used to analyse the competitive landscape as a whole and to provide suggestions on how Cascades could gain a competitive advantage. That being said, the comments on every figure in each of the following subsections are noncomprehensive and nonexhaustive. Their job is to merely provide examples on how the results could be used by Cascades to analyse their competitive landscape.

5.1 Cascades

Cascades (`@CascadesSD`) present themselves on twitter as offering sustainable and innovative solutions for packaging, hygiene and recovery needs (#sourceofpossibilities). They are based out of Kingsey Falls (Quebec) and joined twitter in January 2012. The tweets we were able to collect for `@CascadesSD` span dates from January 2012 till November 2022.

5.1.1 Exploratory Data Analysis

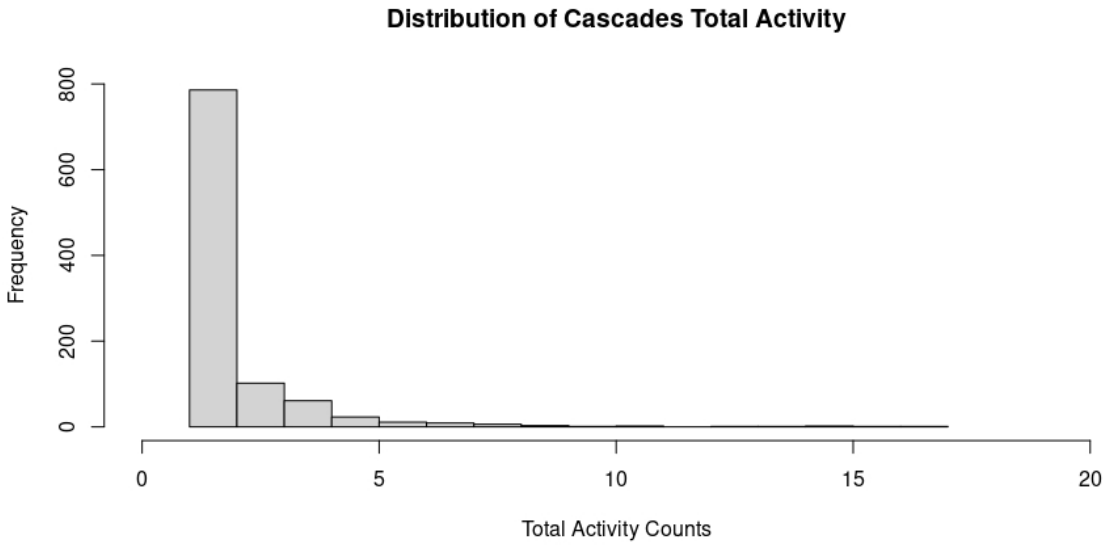


Figure 33: Distribution of total activity counts for the different tweets by Cascades

Looking at the distribution of total activity counts for Cascades, it is clearly non-normal and heavily skewed to the right per figure 33 above. Hence, finding the mean of the total activity counts and taking tweets with total activity values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data would eliminate important information we are looking for and would prevent us from reaching our objective of discovering the topics of the tweets that were voted as most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate ones that were perceived as unimportant by the twitter crowd. This is justified because the total activity data represents discrete counts. Knowing that the 3^{rd} quartile value of total activity count is 2 for Cascades, we will take all the tweets whose total activity is larger than 2. This brings down the total number of Cascades tweets from 1,010 to 224 tweets for analysis.

5.1.2 Identification and Classification of Spikes

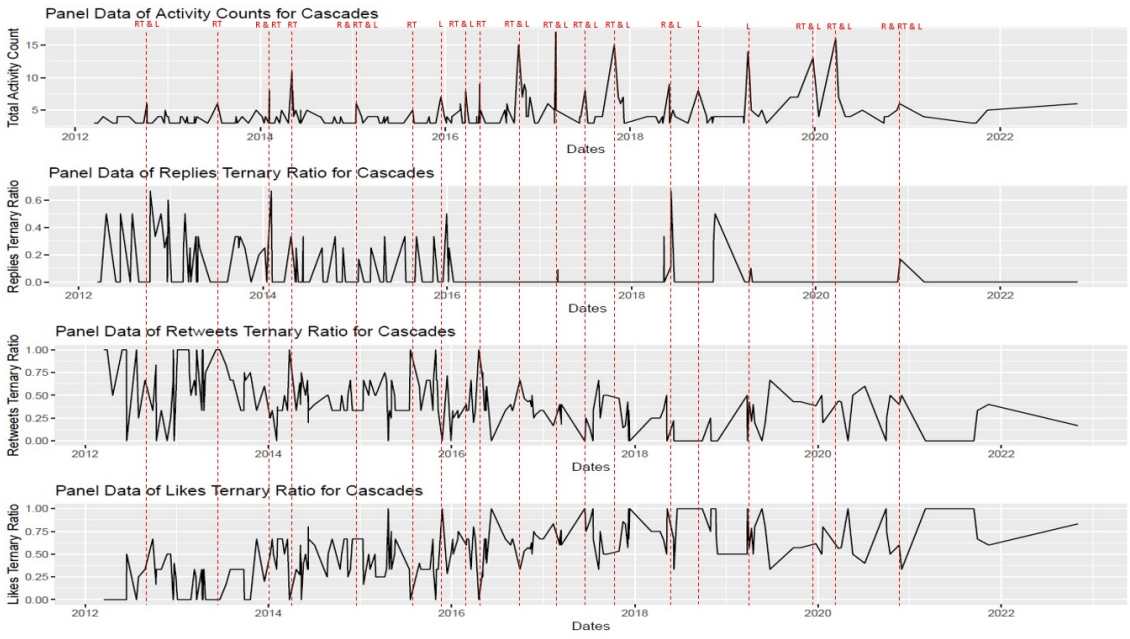


Figure 34: Graphs showing spikes in total activity for Cascades' tweets to correlate with the spikes in ternary ratios of the different activity types. "R" in the graphs represents replies, whereas "RT" represents retweets and "L" represents likes

As can be seen from figure 34 above, Cascades was generally able to keep the twitter crowd engaged over the whole period of analysis. According to the total activity count graph, the twitter crowd found that Cascades started tweeting about hotter topics around the end of the year 2016. Correlating all 4 graphs together, the majority of significantly "spiky" tweets attracted a combination of retweets and likes with very few tweets attracting a combination of all 3 reactions. We also notice the presence of 3 tweets strictly attracting retweets as well as 3 tweets strictly attracting likes. But the tweets with the highest activity counts attracted a combination of retweets and likes. This overview of engagements tells us that Cascades' tweets are potentially seen as less controversial and more credible by the twitter crowd leading to their endorsement in the form of retweets and/or likes. In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down to identify what stimulated the twitter crowd. The next section will help us understand which words attracted each type of reaction over the whole period of analysis.

5.1.3 Frequency of Most Common Words per Reaction

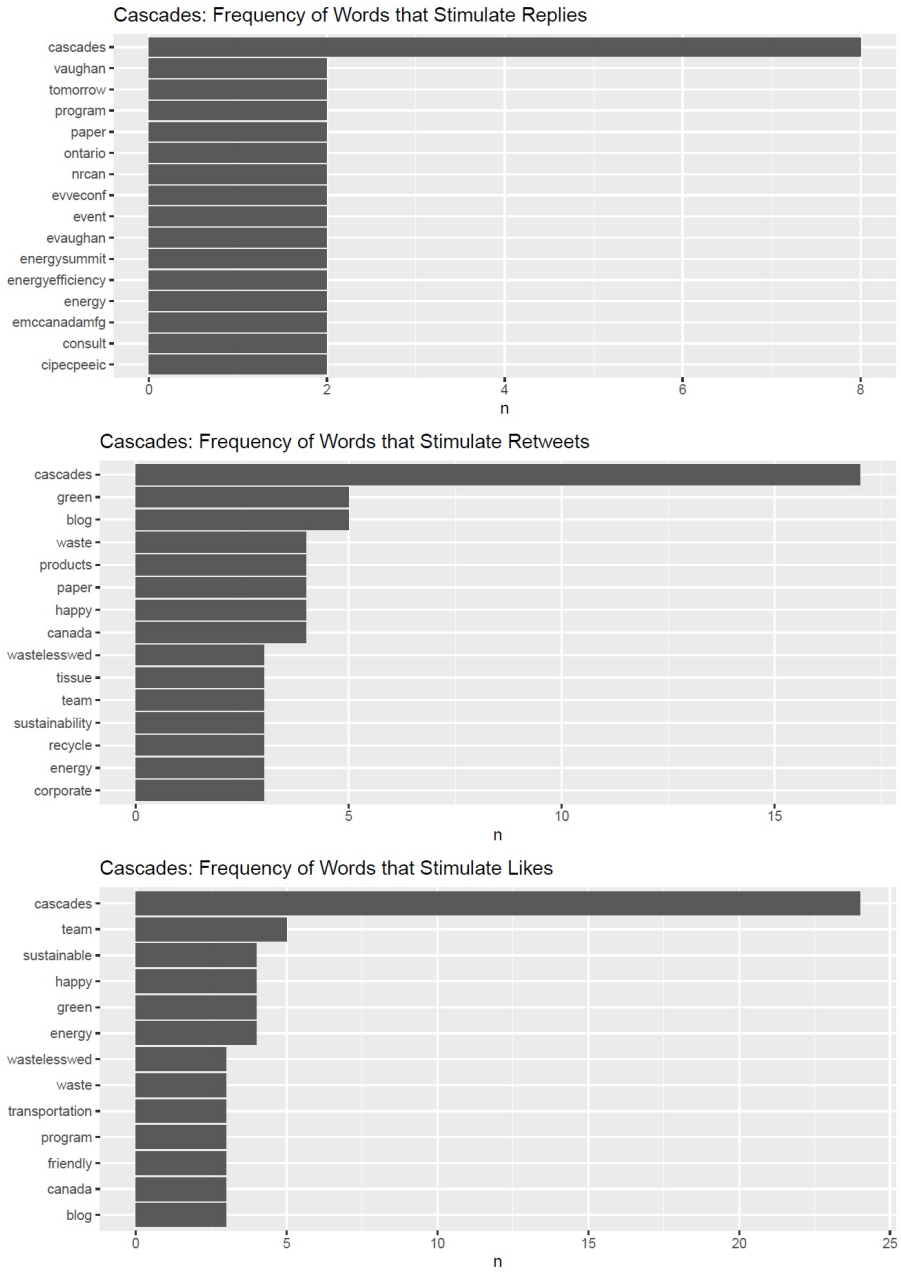


Figure 35: Frequency of words in Cascades' tweets on hot topics

As shown in figure 35 above, the word Cascades is present in all 3 types of reactions and has the highest frequency of all. This is expected as Cascades markets their name in the twitter space. We can also see that Vaughan, Ontario, energy, event, energy efficiency along with energy summit were most repeated in tweets attracting replies. Coupled with the fact from the previous section that very few tweets attracted replies, their word combination indicates that the twitter crowd was motivated to share their thoughts on Cascades attending an energy summit held in Vaughan (Ontario). Whereas the tweets attracting more retweets and likes led

the themes of sustainability and environmental friendliness. These bar charts provide a useful guide for decision makers to implement strategies that align with the wants and needs of the customers and shareholders.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our next point below.

5.1.4 Usage of Most Common Words Over Time

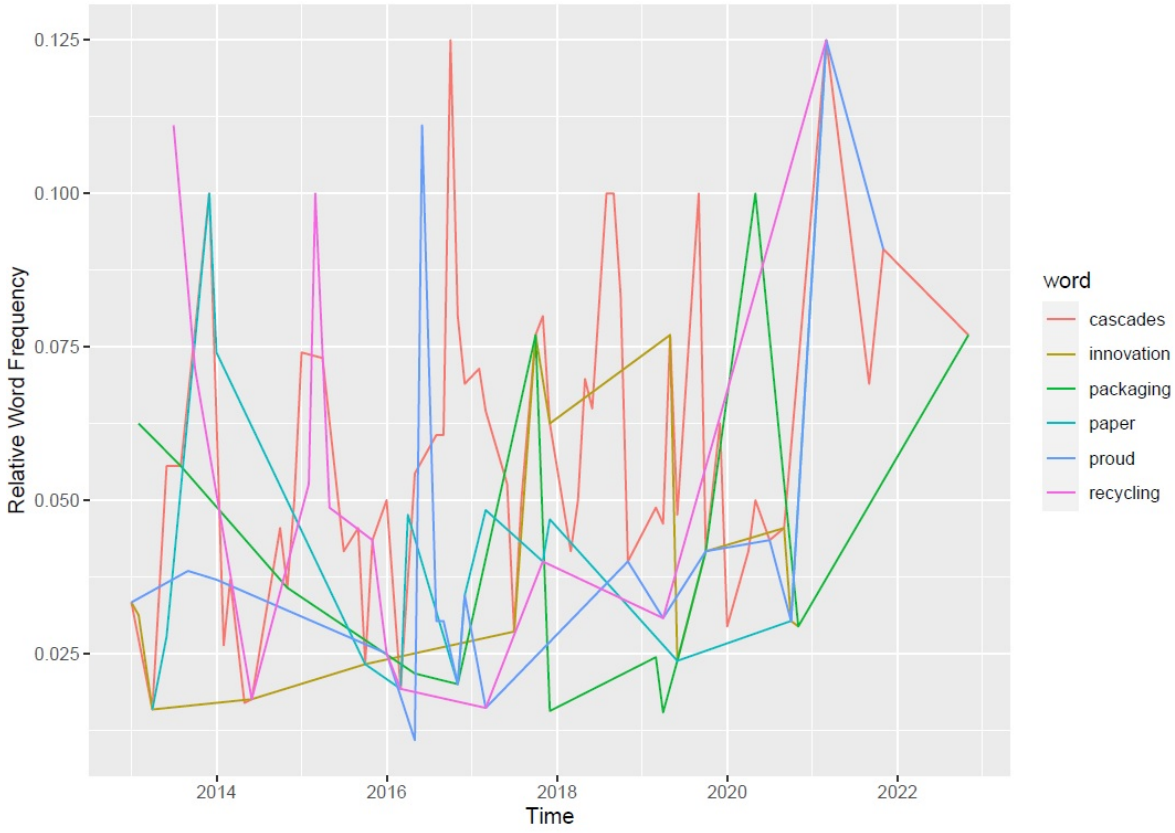


Figure 36: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

Figure 36 above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this

measure would help us perform equitable analysis for all the different twitter accounts.

This figure is useful to highlight the change in Cascades' focus in their public twitter messages over time, which could then be correlated with the previous section to see if their communication focus is actually aligned with what the twitter crowd considers as a hot topic.

5.1.5 Structural Topic Modelling

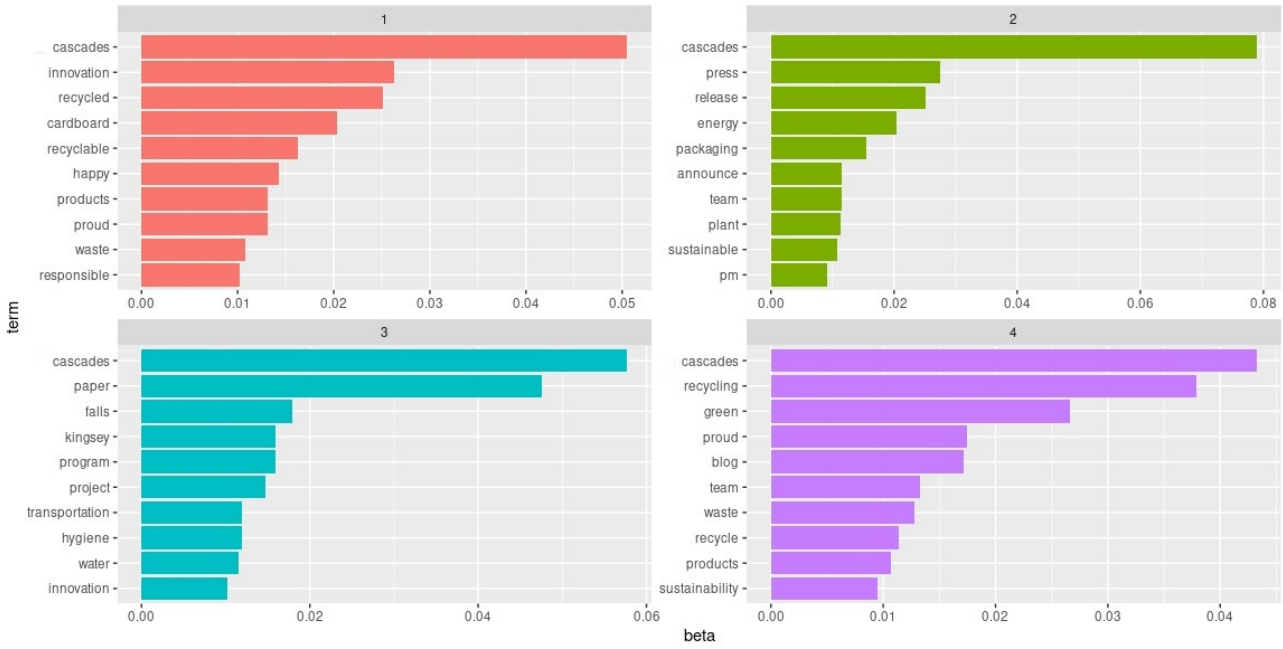


Figure 37: The optimum topic-word combination of the structural topic modelling for Cascades' tweets is 10 words and 4 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents what Cascades focuses on in their business, and thus topic 1 would represent Cascades' business model. As for topic 2, it presents what they usually publish in the press releases and therefore it would represent their traditional media communication with the public. Regarding topic 3, the words shown imply that it talks about subjects related to Cascades' operating model. And the fourth topic discusses what Cascades talks about in their other communication medium, their blog.

This is useful to paint a summarized picture of what Cascades tweets about by organizing and streamlining the large amounts of data into quick bar charts for decision makers. This data could then be used to compare with the topics that competitors tweet about in order for Cascades to see where they succeed and where they lack. Ultimately, this would line them up to launch counter strategies.

Summing up, Cascades tweets about 4 topics: Their business model, their press releases, their operations and their blog.

5.1.6 Sentiment Analysis

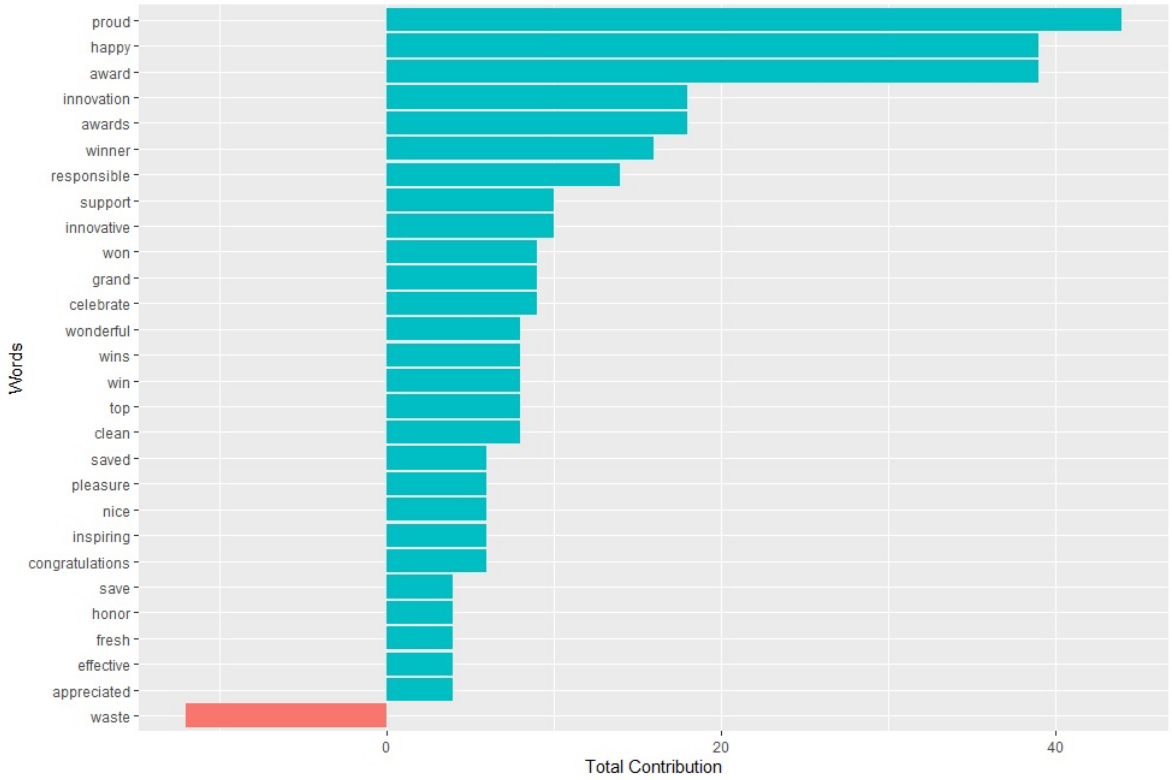


Figure 38: Sentiment analysis for Cascades' qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Cascades. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect Cascades to use positive words when they publish news about the 4 topics we presented in the previous section in order to attract more customers along with more shareholders. This is clearly proven by the top 5 unique words that have the highest contribution to Cascades' overall message sentiment, namely: proud, happy, award, innovation and winner. We can also see that the word waste shows up as negatively affecting Cascades' overall message sentiment. But in reality, a correlation with the previous subsections indicates that the word waste was used in the context of Cascades' contribution to the circular economy. Hence, a positive message after all.

Even though a sentiment analysis on Cascades' tweets is useful, we find that running a sentiment analysis on the crowds' replies to a tweet would be more useful for decision makers, especially if the said tweet garners a very large number of replies that would consume lots of time and resources to go through and analyse manually.

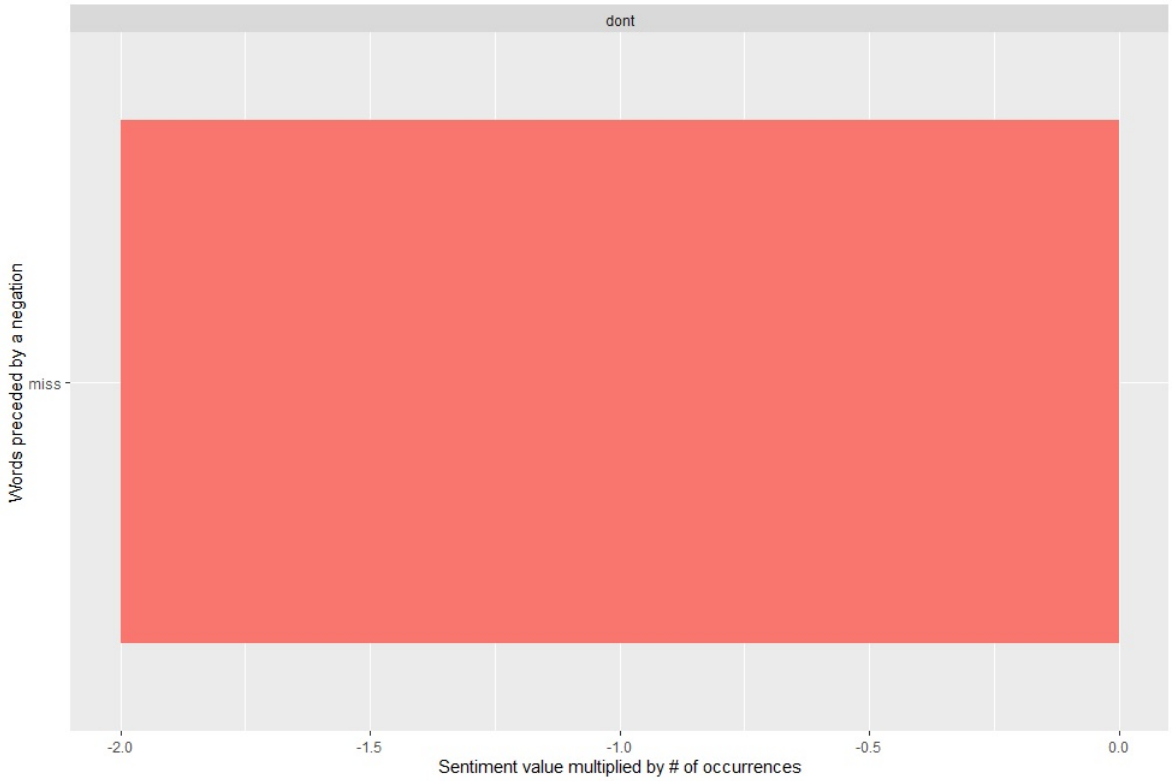


Figure 39: Sentiment analysis for Cascades' tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaluated as the afinn score for each word multiplied by the total number of occurrences.

Our sentiment analysis also showed the presence of 1 negated word present in tweet(s): "miss". Since the word "miss" has a negative connotation, the presence of "don't" before it flips the connotation to positive. As proven by a further drill down, the tweet that included "don't miss" was talking about how Cascades is proud to support more than 25 Quebec athletes in the development of their athletic careers and urges the twitter crowd not to miss their upcoming capsules that will introduce some of these athletes. This tweet attracted a total of 3 likes.

5.1.7 Overall Network of Words in Tweets

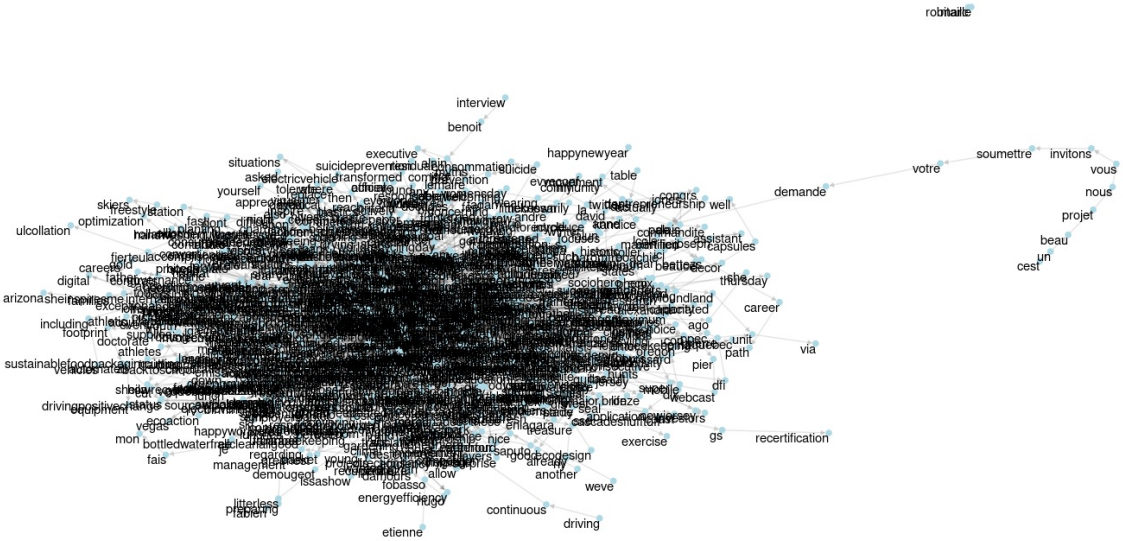


Figure 40: This directed network graph is showing the overall communication strategy of Cascades on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Cascades structures their phrases in their different messages from the beginning to the end. As the darkness of an edge increases, the frequency of the usage of both words connected by that edge increases. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reduce the crowdedness and reveal the words in the dense middle area with highest betweenness centrality. Alternatively, this network graph could be simplified to just include the most common words instead of all the words in Cascades' tweets.

The network graph shown above maps out the current communication strategy of Cascades. This provides a useful guide for decision makers to structure positively perceived twitter communication that appeals to the twitter crowd and to fill up any structural holes in their communication in order to get closer to their targeted audience. It is also interesting to note how the right side of the graph shows a phrase in French that is not part of the dense network in English shown on the other side. This shows that the majority of Cascades' tweets are in English and how the different words are interconnected to tweet about the topics we saw in the previous subsection.

5.1.8 Network of Most Common Used Pairs of Words

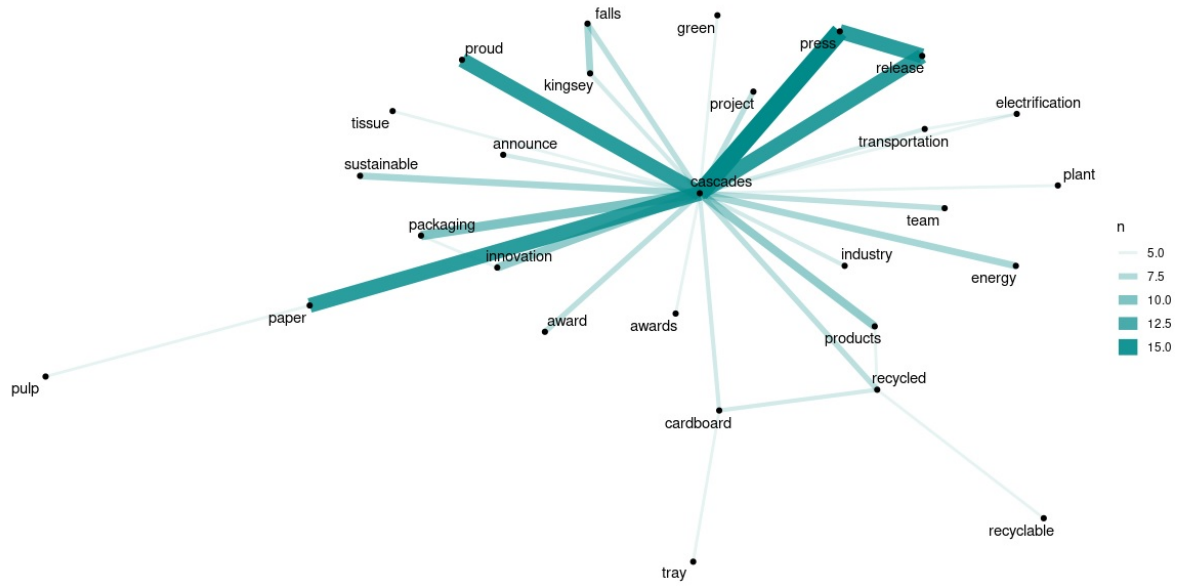


Figure 41: This network graph is showing the most commonly used pairs of words tweeted by Cascades. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

This quasi star-shaped network proves to be advantageous for decision makers seeking to uncover pivot opportunities or new business models. For example, it could indicate Cascades' current innovation strategy which seems to be more concerned with packaging while highlighting what next steps should be taken if decision makers are looking to innovate in their tray products, i.e. Cascades would need to focus on innovating in their cardboard products in order to achieve innovation in the tray products. As for pivot opportunities, this graph helps decision makers to visualize the core components of their current business before utilizing a random network model mentality to combine the different nodes (Barabási 2003). Consequently, uncovering business opportunities. But the best method to uncover pivot opportunities is through carrying out the same network mapping exercise for a different organisation in a different industry and identifying the common words between the different network graphs, before pointing out the structural holes to be bridged between both industries. This methodology proved to be useful in the scientific research world where NASA got inspired by lobsters to create the James Webb Telescope (BBC World Service 2021).

5.2 International Paper

International Paper (`@intlpaperco`) present themselves as transformers of renewable resources into pulp & paper packaging products people depend on everyday. They are based out of Memphis (Tennessee) and joined Twitter in May 2013. The tweets we were able to collect for `@intlpaperco` span dates from July 2013 till November 2022.

5.2.1 Exploratory Data Analysis

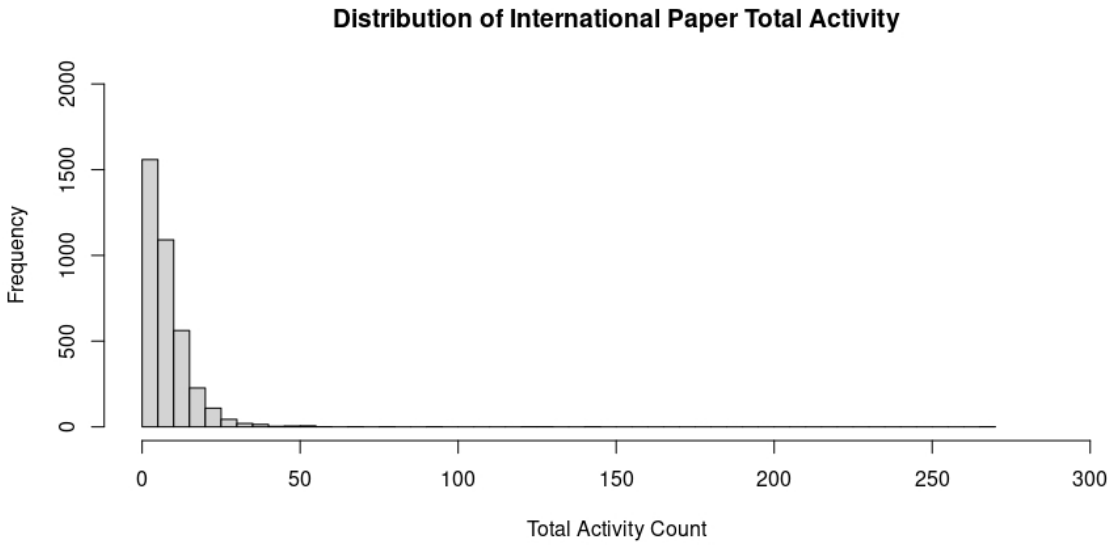


Figure 42: Distribution of total activity counts for the different tweets by International Paper

Looking at the total activity distribution of International Paper, it is non-normal and heavily skewed to the right per the diagram shown above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate International Paper tweets that were perceived as unimportant by the twitter crowd. This is justified because the total activity data represents discrete counts. Knowing that the 3^{rd} quartile value of total activity count is 11 for International Paper, we will take all the tweets whose total activity is larger than 11. This brings down the total number of International Paper tweets from 3,648 to 827 tweets for analysis.

5.2.2 Identification and Classification of Spikes

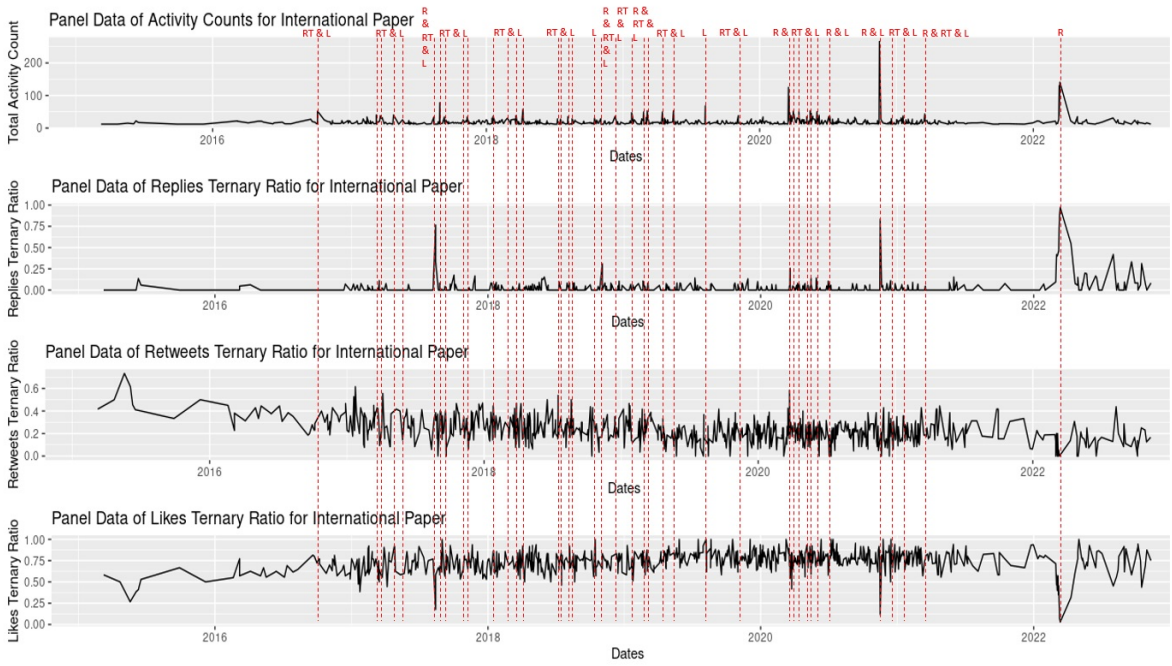


Figure 43: Graphs showing spikes in total activity for International Paper's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes

As can be seen from figure 43 above, International Paper kept the twitter crowd engaged at the same level on average with combinations of reactions over our period of analyses. But the twitter crowd found that International Paper talked about relatively much hotter topics during the last third of our period of analyses, hence the 3 major spikes shown. While 2 of these spikes showed different combinations of reactions, the 3rd spike strictly showed replies which is intriguing. Besides that single tweet that attracted replies only, retweets were always accompanied by likes except for 2 tweets that attracted likes only. This tells us that the word frequency bar charts for retweets will be almost identical to that of likes. The tweet with the highest activity count attracted a combination of replies and likes. This overview of engagements tells us that International Paper's tweets are not always endorsed by the twitter crowd and are sometimes perceived as more controversial, requiring more than just a simple reaction from the twitter crowd, i.e requiring replies.

In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down in order to understand what stimulates their competitor's customers and shareholders. The next section will help us understand which words attracted each type of reaction over the whole period of analysis.

5.2.3 Frequency of Most Common Words per Reaction

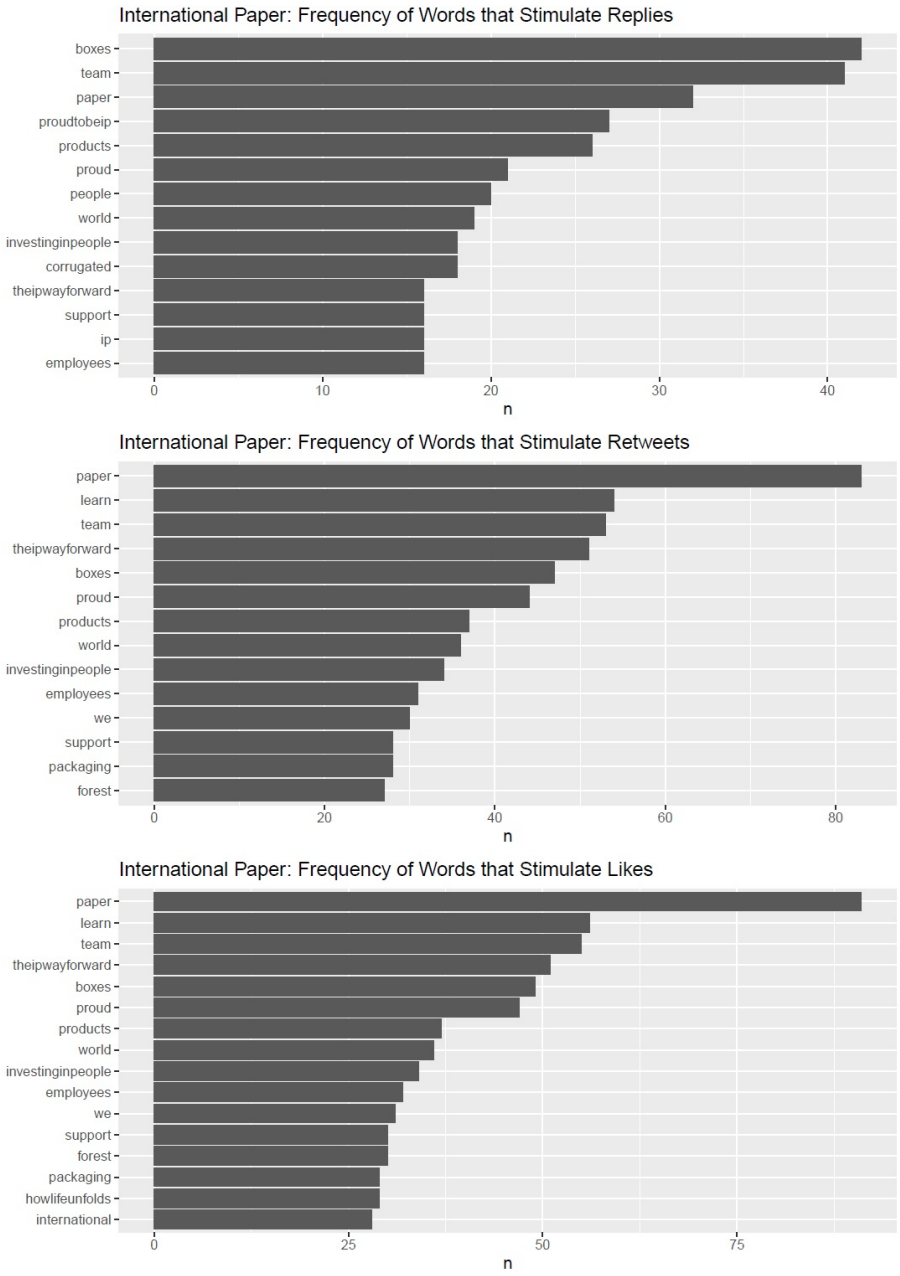


Figure 44: Frequency of words in International Paper's tweets on hot topics

As shown in figure 44 above, the words paper, boxes and team were part of the top 5 words that attracted all 3 types of reactions having the highest frequencies of all. While this is different from Cascades' results, it is expected as International Paper shares a lot in the twitter space about their products and the teams behind their products. Moreover, we can see that the most common words that attracted retweets are almost identical to those that attracted likes which acts as further proof to what we discussed in the previous section. Even though we could also see a lot of most frequent words that are mutually present in all 3

reactions, their frequencies are different in one reaction compared to the other. A good example of that is how International Paper's expression of being "proud" attracts more retweets and likes than replies which indicates that the twitter crowd approves this pride and the achievements behind International Paper's pride. Another example of that is the absence of the word "learn" in the most common words that attract replies and its presence among the other words attracting retweets and likes. This is a further indication that the twitter crowd approves the contents of the tweets that are followed by "learn" more with a hyperlink referring to the tweet topic. On the other hand, International Paper's "corrugated" products attracted more replies than retweets or likes which requires a further drill down to find out why the twitter crowd perceives that as part of a hot topic that motivated them to reply. Consequently, the drill down shows different employees that engage in the corrugated operations of International Paper asking for pay equity during the COVID19 pandemic. Other replies expressed thanks to International Paper for reducing the waste in their corrugated operations. A single reply to the tweets containing the word "corrugated" was asking International Papers to withdraw their operations from Russia in solidarity with Ukraine. One other important point to add is the presence of the word support with boxes at the same time which could indicate that International Paper tweets about their relief efforts and donations in the form of nutritional boxes for food banks. A closer look and correlation could be done with the other subsections that analyse International Paper's tweets.

These bar charts provide a useful guide for Cascades' decision makers to learn more about what triggered replies to International Paper's tweets and why these replies were triggered. This would then help Cascades' strategists to implement tactics that would appeal to the customers and shareholders of International Paper. Cascades could also look at what themes International Paper's twitter crowd endorsed so that they could implement them in order to snatch some of their customers and/or shareholders.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.2.4 Usage of Most Common Words Over Time

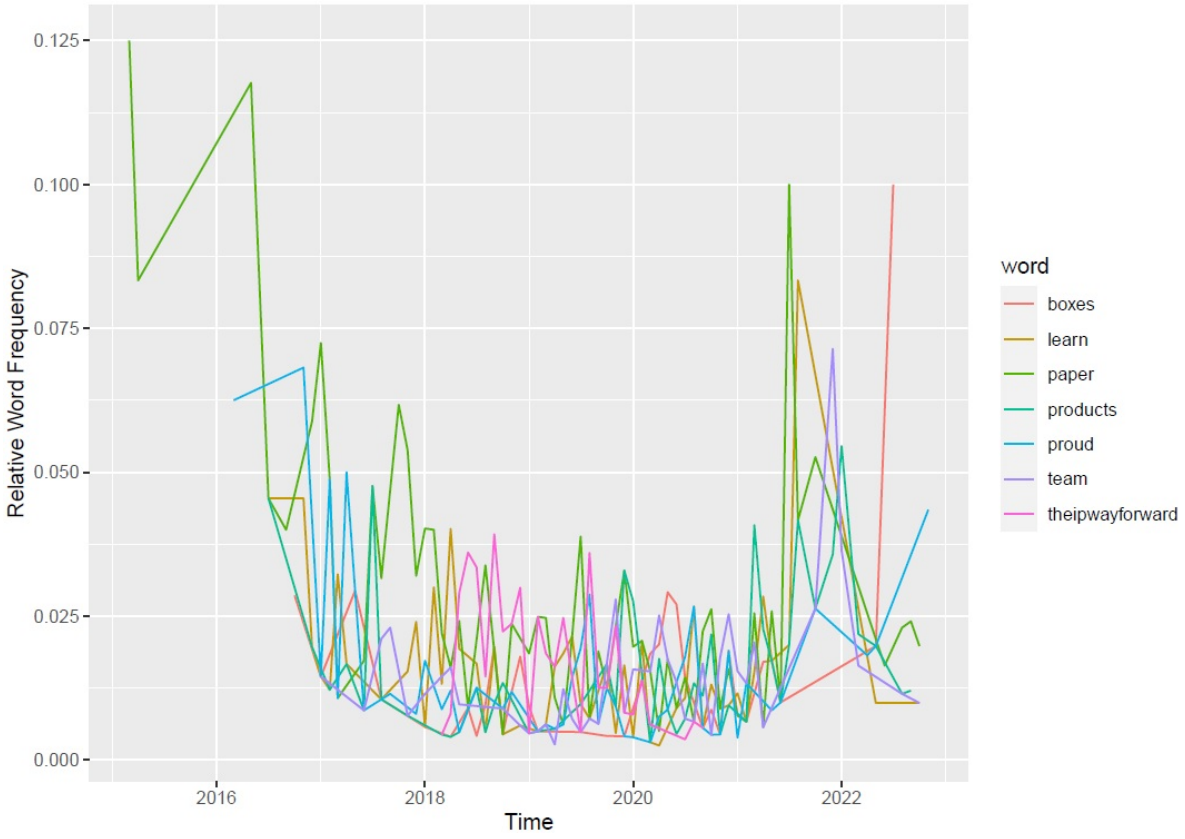


Figure 45: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

Figure 45 above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts.

This figure is useful to highlight the change in International Paper's focus in their public twitter messages over time. Figure 45 is a clear indication to Cascades' team that their competitor, International Paper, increased their focus on their "box" products and hence Cascades should take that into consideration when preparing their next strategic move.

5.2.5 Structural Topic Modelling

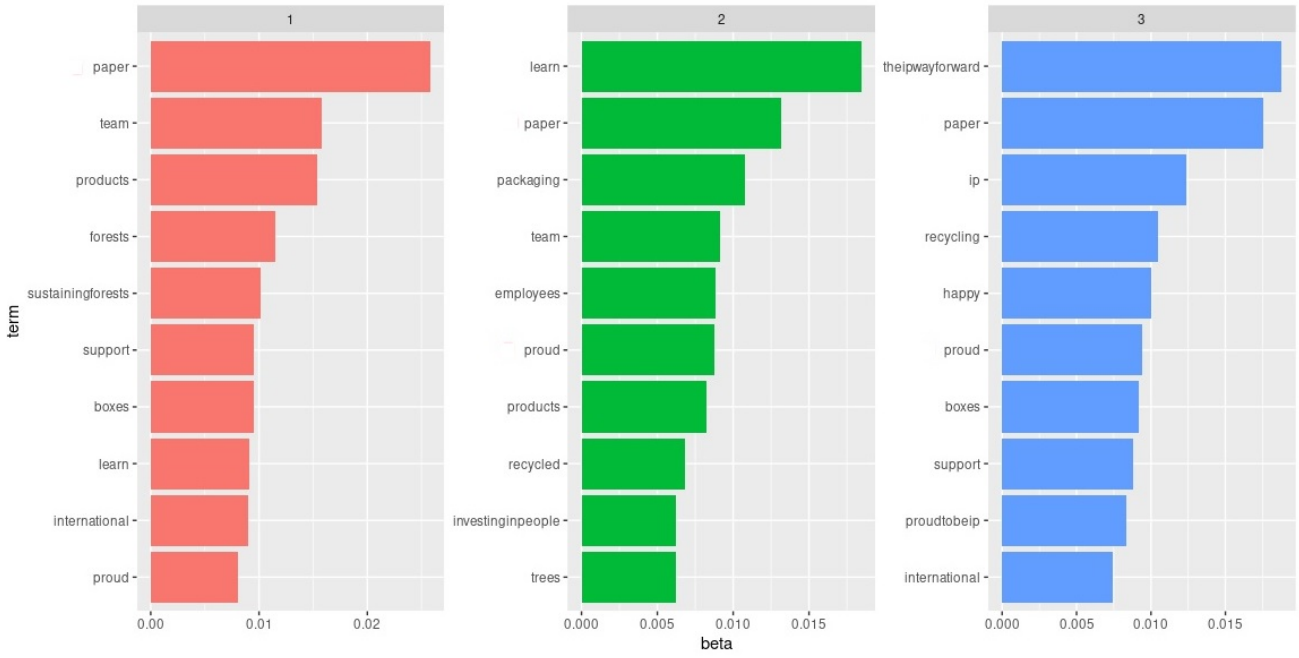


Figure 46: The optimum topic-word combination of the structural topic modelling for International Paper's tweets is 10 words and 3 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents how International Paper's products and operations focus on sustainability towards the environment. As for topic 2, it presents how International Paper invests in the environment, their teams and their communities. Regarding topic 3, the combination of words elude to International Paper's pride moments in the way they do business such as their achievements for example.

There are 2 important points to highlight: Firstly, International Paper refers to themselves either as ip or as International Paper in their tweets. And secondly, International Paper tweets about their international presence. Hence, it would be challenging to discern whether they are simply mentioning their name in the tweets, or are talking about their international presence instead since the capitalization of letters is removed from the most common words in the topic modelling. To remedy that, it is always good to have a larger number of words within the structural topic model in order to get more values of word probabilities and subsequently a clearer view of the topic itself

Summing up, International Paper tweets about 3 topics: Their operations, their impact and their achievements.

5.2.6 Sentiment Analysis

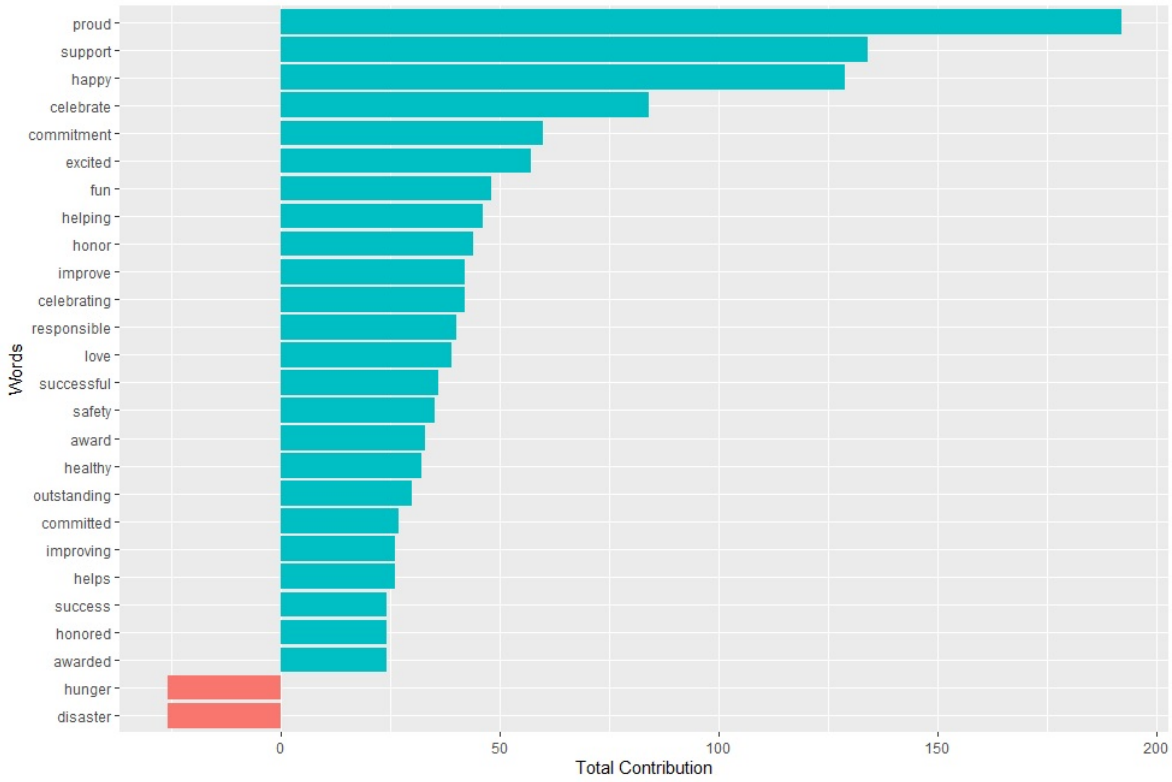


Figure 47: Sentiment analysis for International Paper's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by International Paper. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect International Paper to use positive words when they publish news about the 3 topics we presented in the previous section in order to attract more customers along with more shareholders. This is clearly proven by the top 5 unique words that have the highest contribution to International Paper's overall message sentiment, namely: proud, happy, support, celebrate and commitment. We can also see that the words hunger and disaster show up as negatively affecting International Paper's overall message sentiment. But in reality, a drill down into their tweets containing both of these words indicates that they were used in the context of supporting communities and providing hunger relief or disaster relief. Hence, positive messages after all. In this case, this is useful to inform Cascades' team on how International Paper runs their PR operations.

Furthermore, we find that running a sentiment analysis on the crowds' replies to International Paper's tweets would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

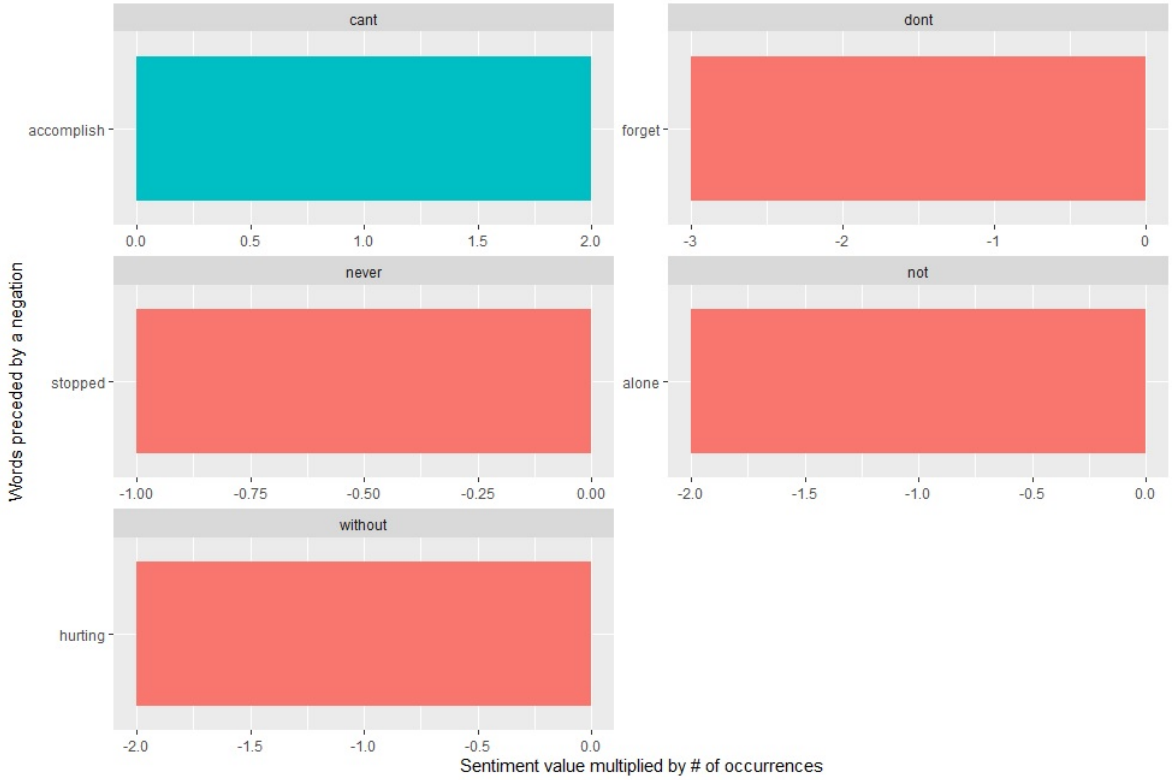


Figure 48: Sentiment analysis for International Paper's tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaludated as the afinn score for each word multiplied by the total number of occurrences.

Our sentiment analysis also showed the presence of 5 negated words present in the tweets, namely: "accomplish", "forget", "alone", "stopped", "hurting". Thus, the presence of negation before these words flips their connotation. Overall, the total afinn sentiment score of these words does not have a sizable effect on the total sentiment score shown in the previous bar chart.

5.2.7 Overall Network of Words in Tweets

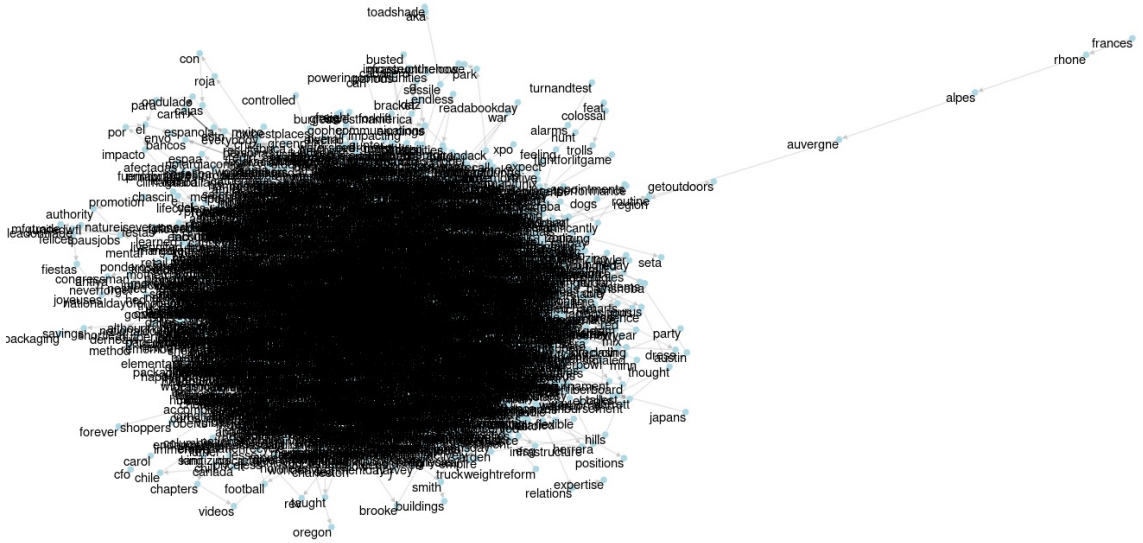


Figure 49: This directed network graph is showing the overall communication strategy of International Paper on Twitter. The labelled nodes represent all the words used and the directed edges indicate how International Paper structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of International Paper. This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd. This also helps Cascades to fill up any structural holes in their communication as well as their strategy. It is also useful to uncover what International Paper talks about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.2.8 Network of Most Common Used Pairs of Words

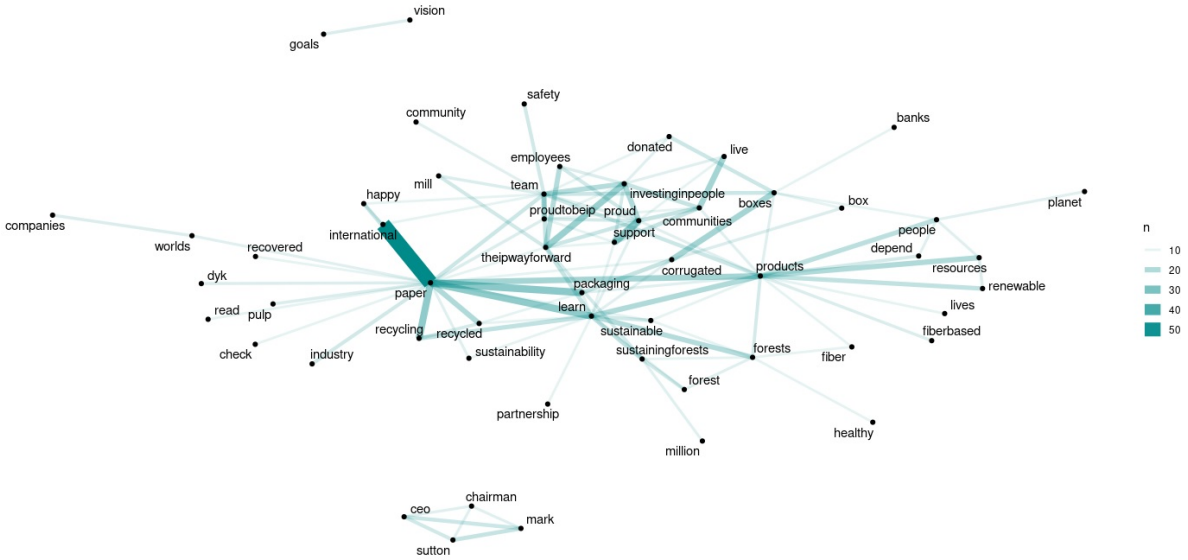


Figure 50: This network graph is showing the most common usage of the pair of words tweeted by International Paper. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that International Paper tweets about. We can clearly see the presence of several clusters and small worlds that indicate topics being discussed (Goyal, Van Der Leij, and Moraga-González 2006; Uzzi and Spiro 2005; Watts and Strogatz 1998). For example, the word "products" forms a quasi-star network which is loosely and directly connected to the word "team" that forms its own quasi-star network. Both of these map out what products International Paper produces and how they operate to produce them. Mapping out both of these aspects provides a clear visual map on what International Paper is doing and how they are doing it in order for Cascades' strategy room to explore where they are lacking. Another interesting aspect of this graph is how it indicated, via the independent complete graph on the bottom left, that the chairman and CEO of International Paper is called Mark Sutton. One final point to note here is, as we expected from the word frequency section, the word "boxes" implies the boxes that International Paper produces as well as the boxes they donate to food banks.

Needless to say, it would be interesting for Cascades to drill down into what other industries or sectors use "fiber" or "fiberbased" products as shown in International Paper's graph on the lower right end. This would help them explore and uncover pivot opportunities in order to grow the market value and their market share.

5.3 Mighty Earth

Mighty Earth (`@standmighty`) present themselves on twitter as a group that runs global campaigns to protect rain forests, make agriculture sustainable and solve climate change (#fightforforests #neverimpossible). They are based out of Washington (District of Columbia) and joined Twitter in August 2016. The tweets we were able to collect for `@standmighty` span dates from August 2016 till November 2022.

5.3.1 Exploratory Data Analysis

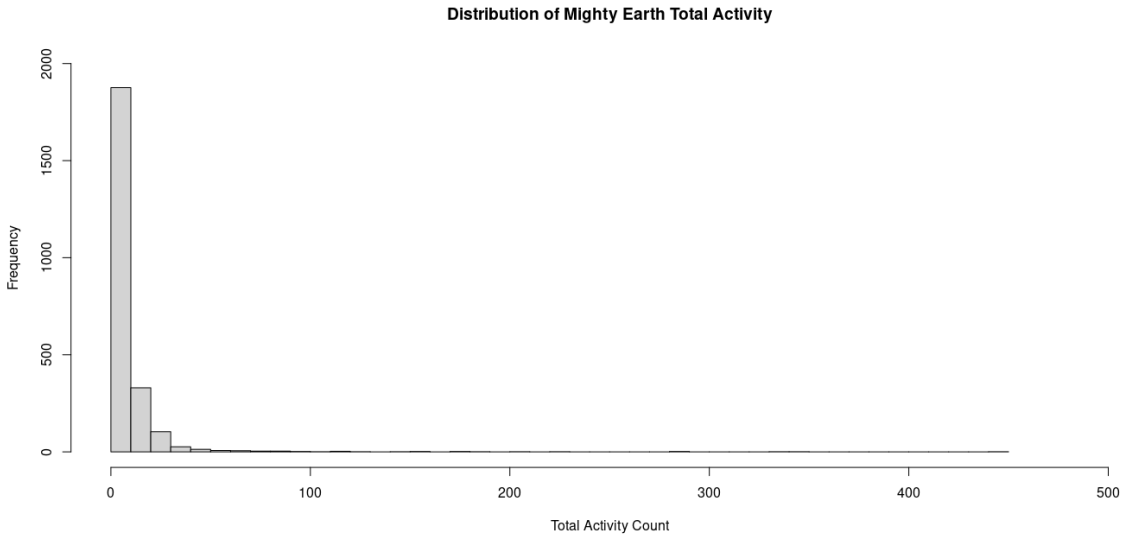


Figure 51: Distribution of total activity counts for the different tweets by Mighty Earth

Looking at the Twitter activity distribution of Mighty Earth, it is non-normal and heavily skewed to the right per the histogram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate Mighty Earth's tweets that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 9 for Mighty Earth, we will take all the tweets whose total activity is larger than 9. This brings down the total number of Mighty Earth tweets from 2,390 to 580 tweets for analysis.

5.3.2 Identification and Classification of Spikes

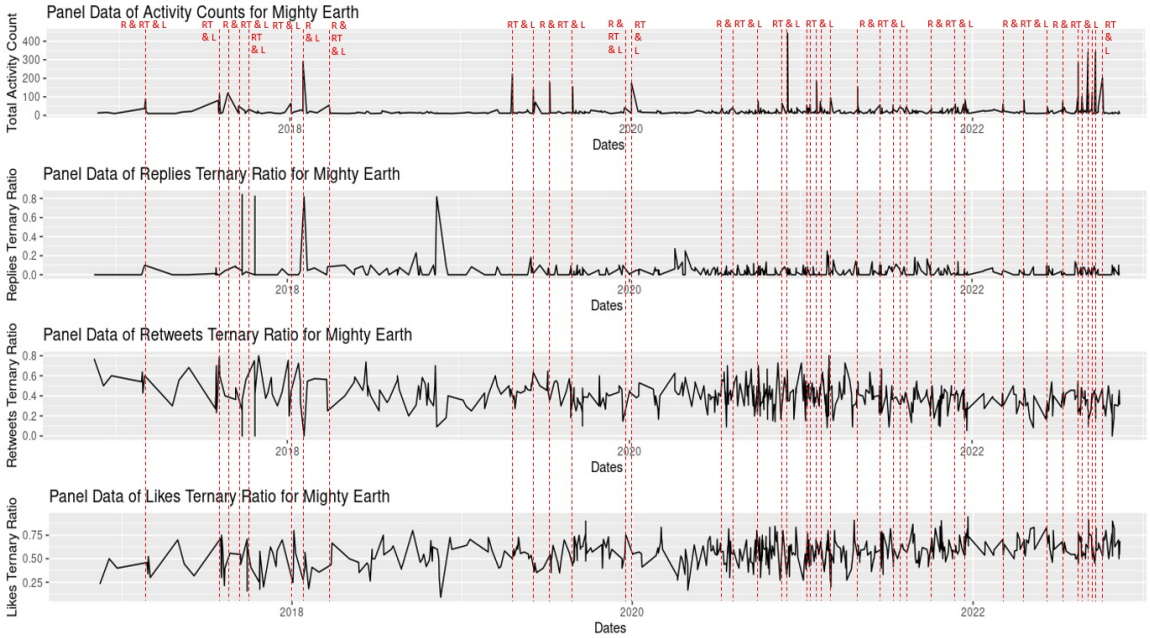


Figure 52: Graphs showing spikes in total activity for Mighty Earth's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Mighty Earth did a better job in keeping the twitter crowd engaged in the second half of the period of analysis compared to the first half. According to the total activity count graph, the twitter crowd found that Mighty Earth started tweeting about hotter topics around the beginning of the year 2020. Another important point to note is the presence of very few significant spikes in the replies ternary ratio graph which indicates that the major reactions to Mighty Earth's "spiky" tweets were just retweets and likes. Correlating all 4 graphs together, all the significantly "spiky" tweets attracted different combinations of replies, retweets and likes with no tweets attracting just one unique reaction. The tweets with the highest activity counts attracted a combination of retweets and likes. This overview of engagements tells us that Mighty Earth's tweets are endorsed by the twitter crowd and are perceived as credible.

In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down to see who is on Mighty Earth's nice list VS. their naughty list. We could expect the 3 word frequency bar charts of Mighty Earth in the next section to contain more or less the same words, albeit with different frequencies. Nevertheless, it is a good starting point to help us understand which words attracted each type of reaction

over the whole period of analysis.

It is important to contextualize Cascades' analysis of Mighty Earth's tweets by keeping in mind that Mighty Earth is an environmental lobby group that influences customers and shareholders alike. In other words, Cascades' strategists should integrate Mighty Earth's criticisms and topics into their strategies in order to avoid any adverse effects to their reputation which may lead to loss in market share.

5.3.3 Frequency of Most Common Words per Reaction

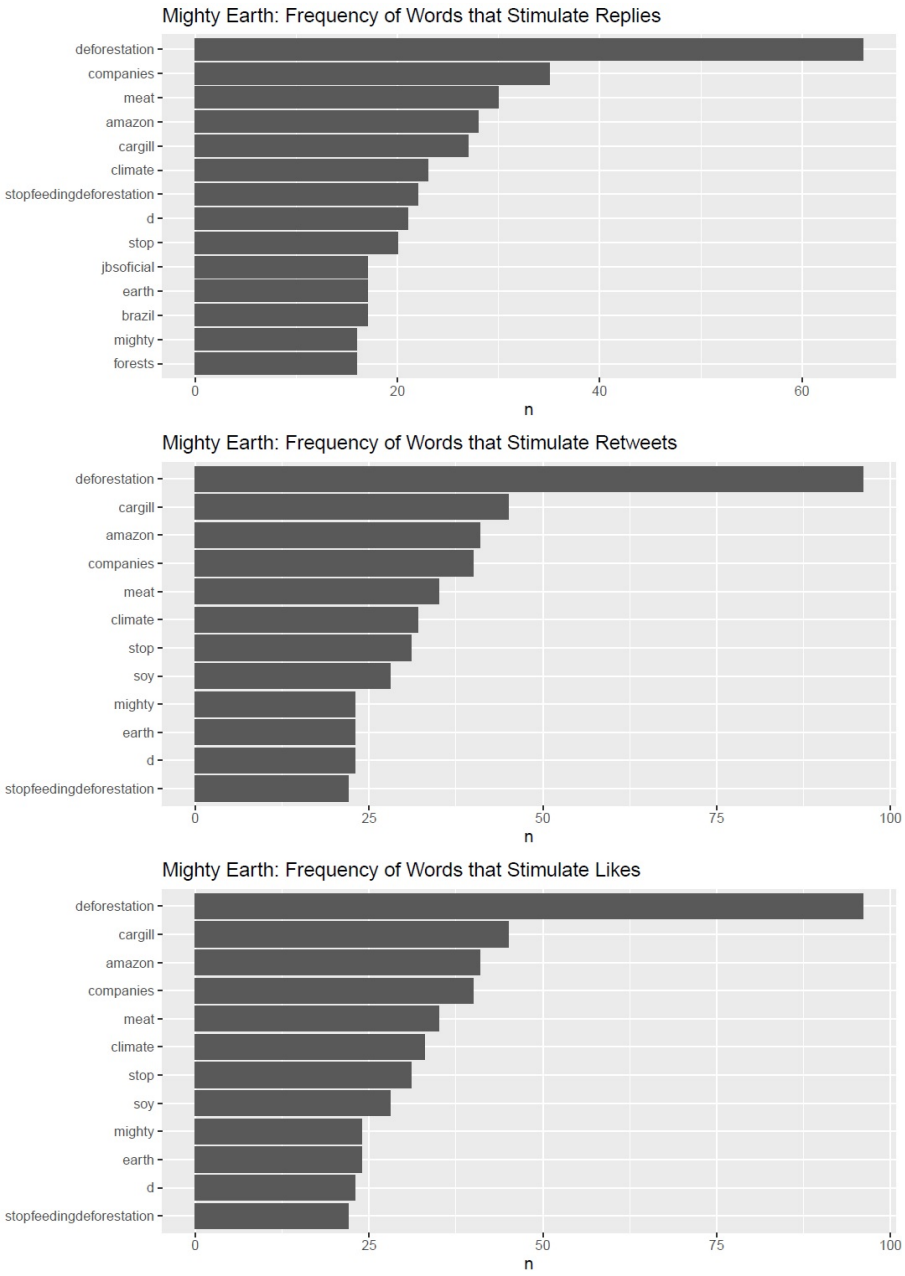


Figure 53: Frequency of words in Mighty Earth's tweets on hot topics

As shown in figure 53 on the previous page, the words that stimulate retweets and the words that stimulate likes are identical. We can also see that the words that stimulate replies are quite close to the words that stimulate the 2 other reactions. We can see that the top 5 most frequent words in Mighty Earth's tweets that attracted the 3 reactions are all the same albeit having different counts, namely: "deforestation", "companies", "meat", "amazon", "cargill" followed by "climate" in 6^{th} place. In this case, Cascades should look closer at the tweets corresponding to these words to extract what they shouldn't be doing in their business. For example, stop any transactions with companies such as Cargill (a meat and poultry company) that may attract bad publicity, or have negative effects on forests in general or the amazon in particular, or even contribute to climate change.

Said differently, these bar charts provide a useful guide for Cascades' decision makers to explore what they should and shouldn't do by focusing on the topics of the influencer's tweets. Consequently, this would then help Cascades' strategists to implement tactics that would appeal to Mighty Earth and their audience.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.3.4 Usage of Most Common Words Over Time

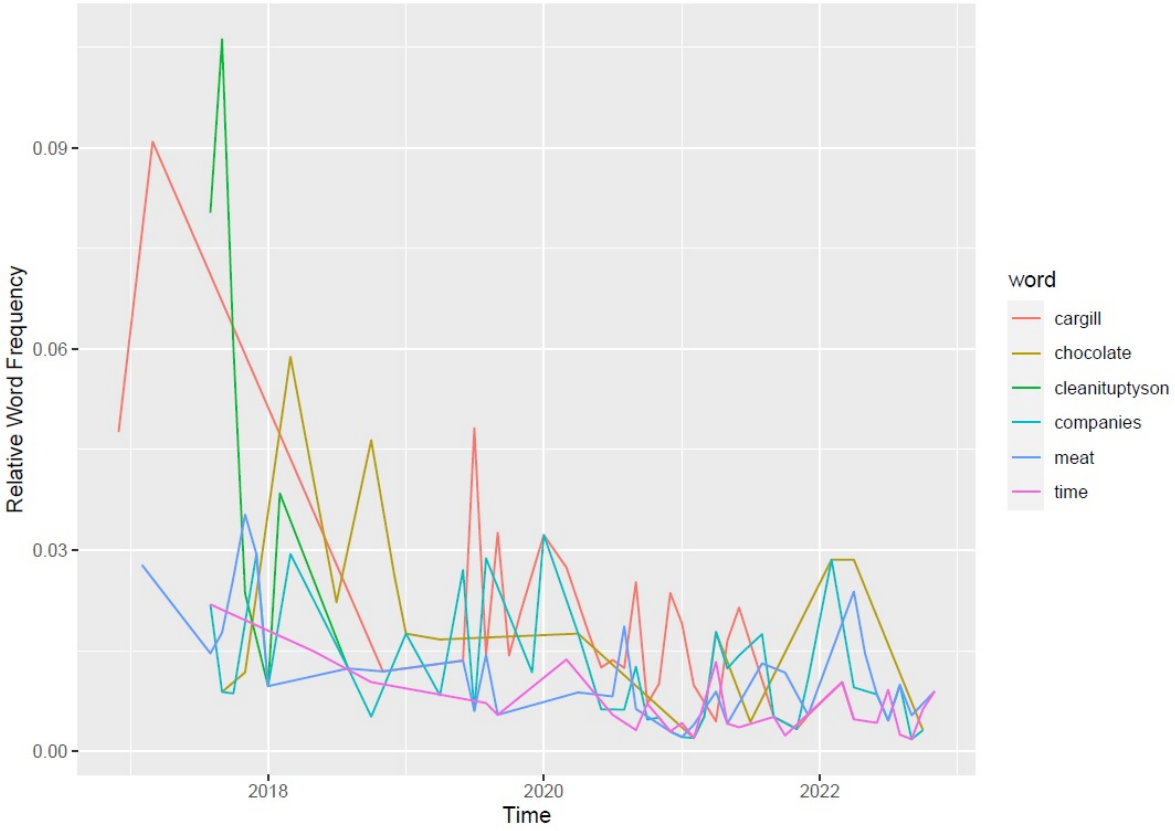


Figure 54: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts.

Figure 54 is useful to highlight the change in Mighty Earth's focus in their public twitter messages over time. We can see here that they stopped tweeting about about Cargill around the end of 2018 before starting to tweet again around mid 2019 until mid 2021. Whereas chocolate, meat and companies were always prominent in Mighty Earth's tweets over the whole period of analysis. This figure could be used by Cascades to track the topics that Mighty Earth uses to influence customers and shareholders in order to realign their offerings

with the interests of those stakeholders.

5.3.5 Structural Topic Modelling

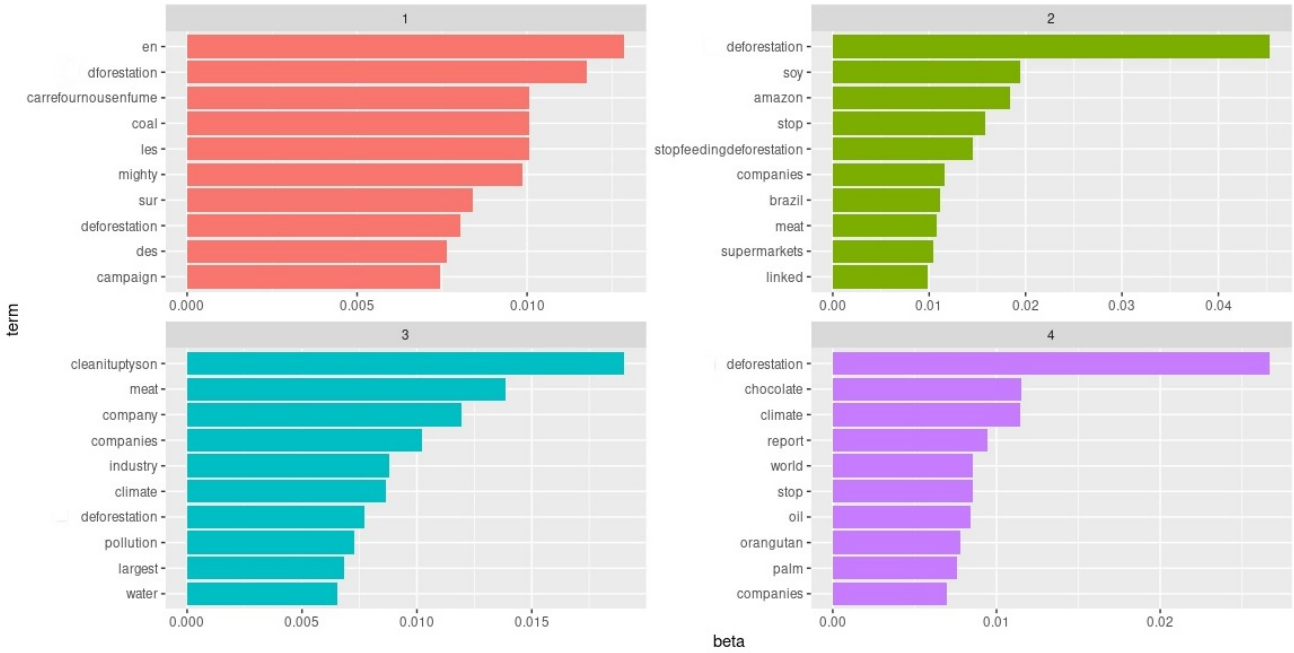


Figure 55: The optimum topic-word combination of the structural topic modelling of Mighty Earth's tweets is 10 words and 4 topics

We can clearly see that Mighty Earth is trying to fight deforestation in all 4 topics. But drilling down further into these topics and based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents how Mighty Earth exposes the damage made to the environment through the extraction of coal. As for topic 2, it presents how Mighty Earth exposes the damage that meat and soy products made to the environment with a focus on the Amazons and Brazil. Regarding topic 3, the combination of words elude to the effects of the different non-sustainable practices made by companies. And finally topic 4 talks about how the extraction of cocoa and palm oil is affecting the forests.

It is important to note that Mighty Earth is an organisation whose activity on twitter focuses on protecting the earth by lobbying for or against organisations that are or are not environmentally-friendly. In other words, Mighty Earth's goal is to protect the earth. That being said, it would be challenging in this case to discern from the structural topic modelling whether they are referring to their name, Mighty Earth, in the tweet or are referring to the planet we live in when the word "earth" comes up. To remedy that, it is always good to have a larger number of words within the structural topic model in order to get more values of word probabilities and subsequently a clearer view of the topic itself.

Summing up, Mighty Earth tweets about 4 non-environmentally friendly topics: Coal, soy and

meat, pollution in addition to palm oil and cocoa.

5.3.6 Sentiment Analysis

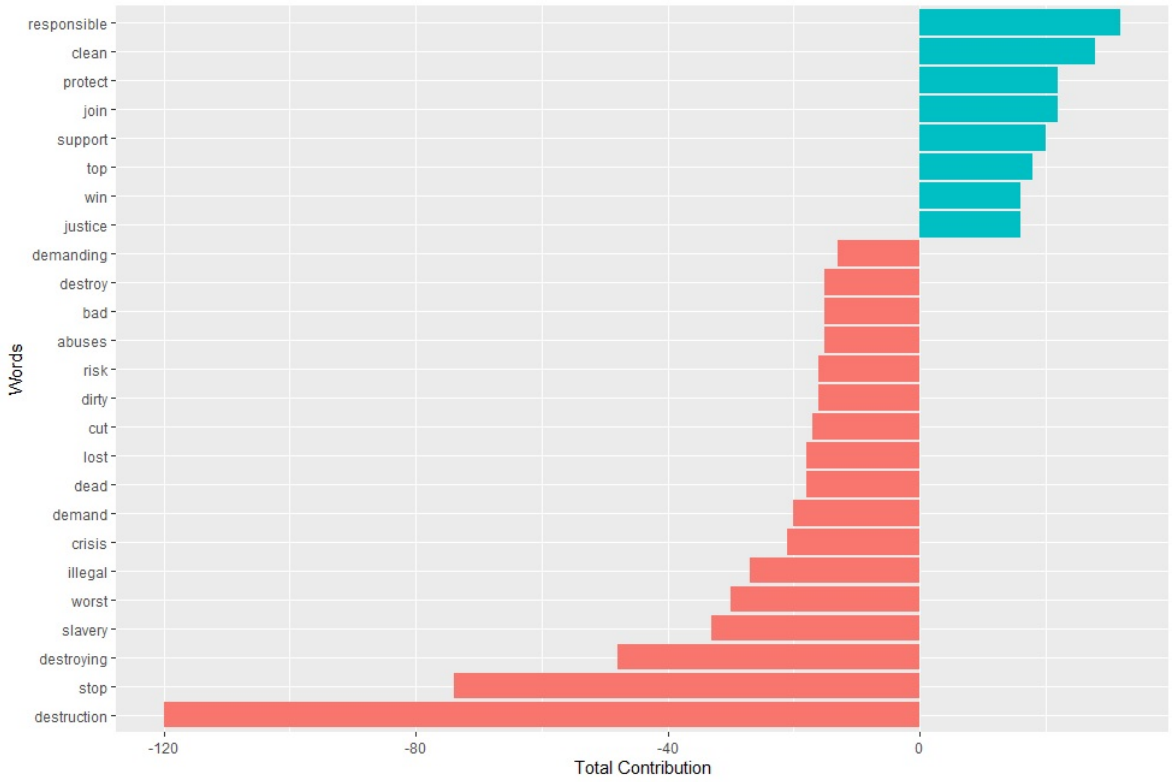


Figure 56: Sentiment analysis for Mighty Earth's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Mighty Earth. As a lobbying group that has the power to influence the twitter audience through criticizing practices that are not environmentally friendly, we would expect Mighty Earth's tweets to mostly use negative words when they publish news about the 4 topics we presented in the previous section. This is clearly proven by the top 5 unique words that have the highest contribution to Mighty Earth's overall message sentiment, namely: destruction, stop, worst, slavery and illegal. We can also see that the words "responsible", "clean", "protect", "join" and "justice" show up as positively affecting Mighty Earth's overall message sentiment. But in reality, a drill down into their tweets containing these words indicates that they were defining who is "responsible" for deforestation, asking a company to "clean" up their pollution footprint, urging the twitter audience to "join" the move to "protect" and bring "justice" to

the amazon in Brazil. Hence, these words were being used to relay a negative context.

Just like the previous subsections, this could be used by Cascades to identify the most negative topics, correlate them with the activity counts to identify what the twitter audience voted for as the worst malpractice. But we also find that running a sentiment analysis on the crowds' replies to Mighty Earth's tweets would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what malpractices should be corrected or avoided with the limited resources available.

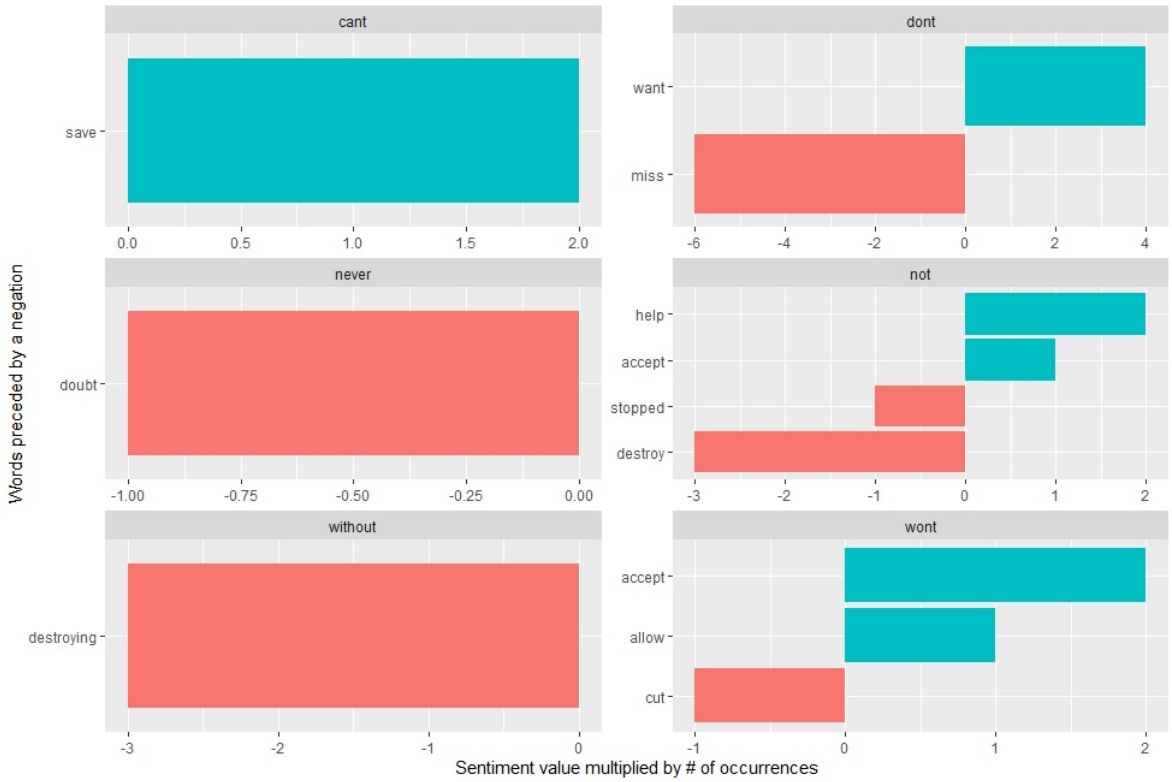


Figure 57: Sentiment analysis for Mighty Earth's tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaludated as the afinn score for each word multiplied by the total number of occurrences.

Our sentiment analysis also showed the presence of 12 negated words present in the tweets. Thus, the presence of negation before these words flips their connotation. Overall, the total afinn sentiment score of these words does not have a sizable effect on the total sentiment score shown in the previous bar chart.

5.3.7 Overall Network of Words in Tweets

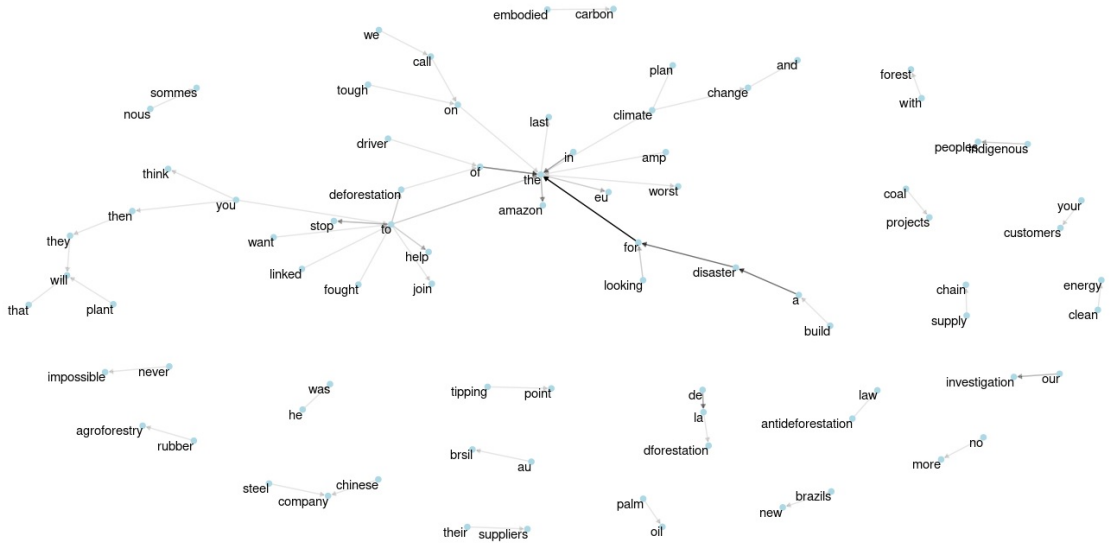


Figure 58: This directed network graph is showing the overall communication strategy of Mighty Earth on Twitter. The labelled nodes represent the most repeated words used and the directed edges indicate how Mighty Earth structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. This overall graph was simplified for Mighty Earth in order to highlight what they talk about the most as an environmental lobby group.

The network graph shown above maps out the current communication strategy of Mighty Earth. This provides a useful guide for Cascades' decision makers to structure environmentally-friendly tweets that appeal to the twitter crowd. What is interesting about this graph is that it summarizes the different environmental offenders into independent networks that surround the interconnected network showing some of the adverse environmental effects. This also helps Cascades to fill up any structural holes in their communication and to uncover what and how Mighty Earth criticizes environmental crimes, which in turn could be used to get closer to customers and shareholders.

5.3.8 Network of Most Common Used Pairs of Words

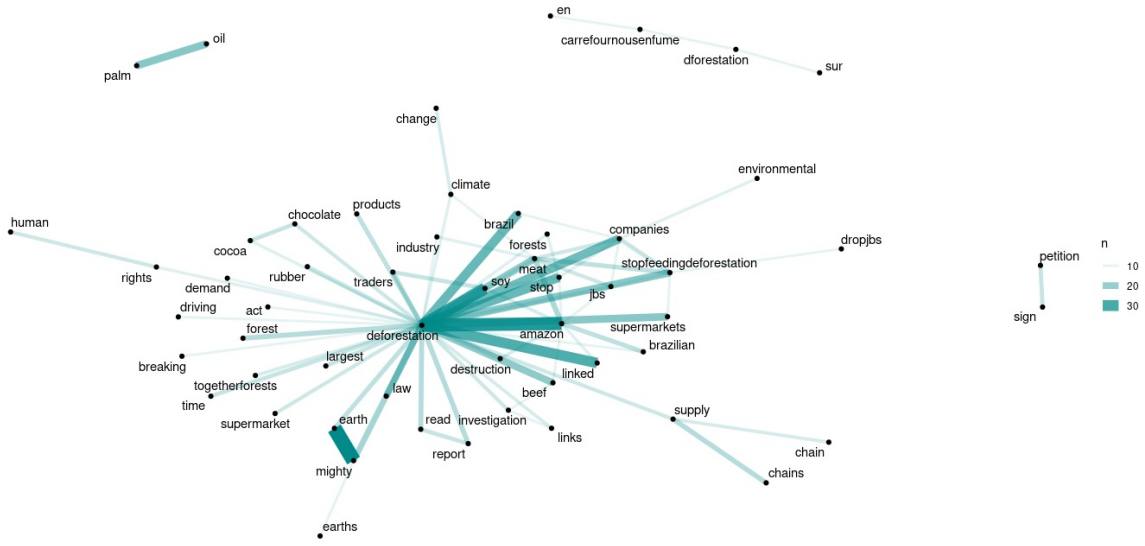


Figure 59: This network graph is showing the most common usage of the pair of words tweeted by Mighty Earth. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that Mighty Earth tweets about. We can clearly see the presence of a quasi star shaped graph with the word "deforestation" having the highest betweenness, degree and eigenvector centralities. This shows how important the concept of deforestation is to Mighty Earth. We can also see that the destruction of the forests in the amazon and the effect of meat and soy production on the environment on the right side of the network were more frequent compared to the topics on the left side of the network.

It would be useful for Cascades to either not engage in the topics presented in the graph above, or to align their engagement with how Mighty Earth presents these topics in order to maintain their image and reputation.

5.4 Resolute Forest Products

Resolute Forest Products (`@resolutefp`) present themselves on twitter as a a leading producer of a diverse range of wood, pulp, tissue and paper products, which are marketed in 60 countries. They are based out of Montreal (Quebec) and joined Twitter in September 2011. The tweets we were able to collect for `@resolutefp` span dates from June 2012 till November 2022.

5.4.1 Exploratory Data Analysis

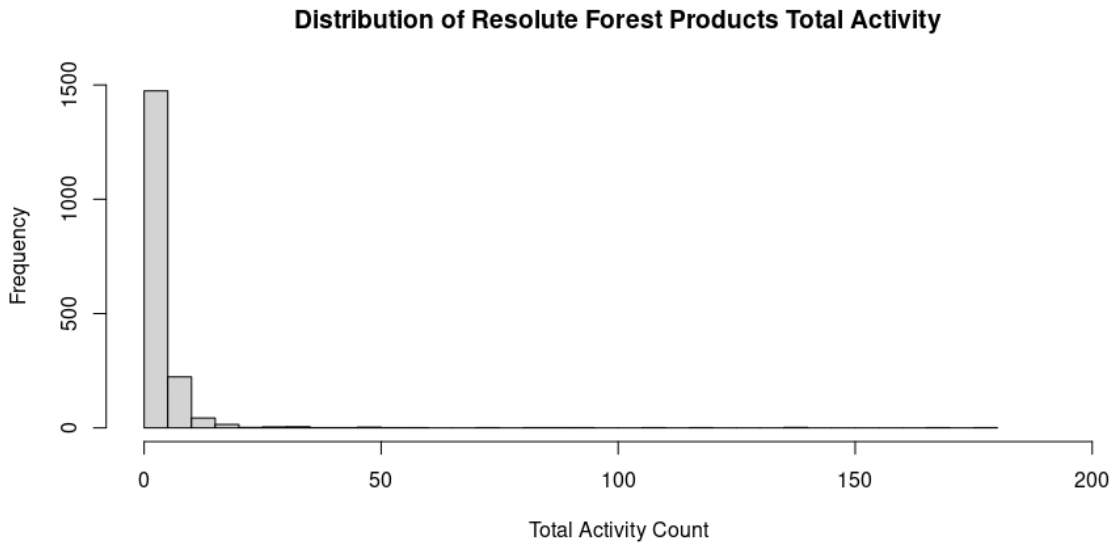


Figure 60: Distribution of total activity counts for the different tweets by Resolute Forest Products

Looking at the Twitter activity distribution of Resolute Forest Products, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by Resolute Forest Products that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 4 for Resolute Forest Products, we will take all the tweets whose total activity is larger than 4. This brings down the total number of tweets by Resolute Forest Products from

1,783 to 434 tweets for analysis.

5.4.2 Identification and Classification of Spikes

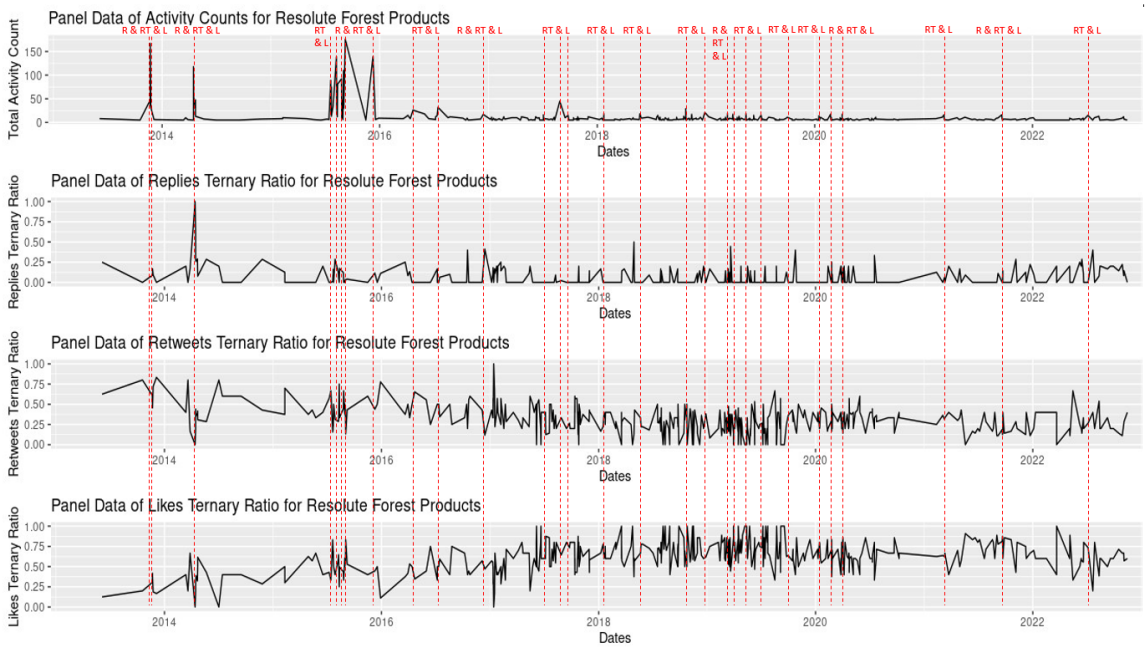


Figure 61: Graphs showing spikes in total activity for Resolute Forest Product's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Resolute Forest Products did a better job in keeping the twitter crowd engaged in the second half of the period of analysis compared to the first half. But according to the total activity count graph, the twitter crowd found that Resolute Forest Products tweeted about hotter topics during the first half of the period of analysis. Hence, the presence of significant spikes between the end of the year 2013 and the year 2016. Another important point to note is the presence of just one significant spike in the replies ternary ratio graph which indicates that the majority of reactions to Resolute Forest Products' "spiky" tweets were just retweets and likes. Correlating all 4 graphs together, all the "spiky" tweets attracted different combinations of replies, retweets and likes with a relatively low concentration of replies and no tweets attracting a just a single reaction. The tweets with the highest activity counts attracted a combination of replies, retweets and likes. This overview of engagements tells us that the tweets made by Resolute Forest Products are not always endorsed by the twitter crowd or viewed as credible which motivates them to react with more than just a retweet or like, i.e reply.

In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down. The next subsection will help us understand which words attracted each type of reaction over the whole period of analysis.

5.4.3 Frequency of Most Common Words per Reaction

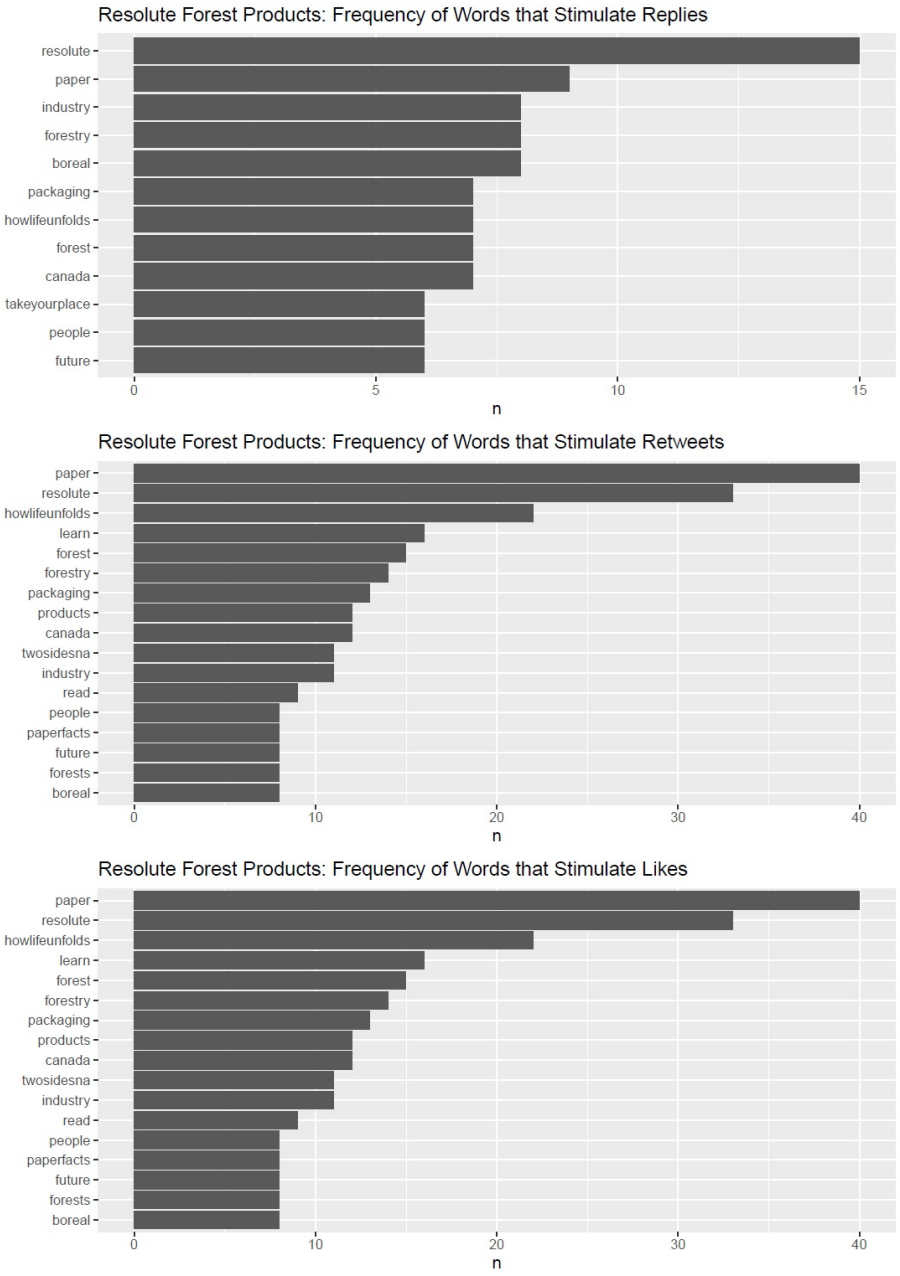


Figure 62: Frequency of words in the tweets of Resolute Forest Products on hot topics

As shown in the figure above, the words resolute, paper and forestry were part of the top 6 words that attracted all 3 types of reactions having the highest frequencies of all. While this is slightly different from Cascades' results, it is expected as Resolute Forest Products markets

their name, their products and how they do their business. Moreover, we can see that the most common words that attracted retweets are identical to those that attracted likes which acts as further proof to what we discussed in the previous section. Even though we could also see a lot of most frequent words that are mutually present in all 3 reactions, their frequencies are different in one reaction compared to the other. For example even though the word "boreal" shows up in all 3 bar charts, it does not show up as part of the top 5 or even top 10 most frequent words attracting retweets and likes contrary to replies. A further drill down into the tweets containing the word "boreal" indicates that Resolute Forest Products was receiving harsh or constructive feedback on their practices in the Canadian Boreal forest. This in turn triggered Resolute Forest Products to repeatedly emphasize their commitment to sustainable forestry in the Boreals. Hence, the appearance of the word "boreal" as part of the top 5 words attracting replies.

These bar charts provide a useful guide for Cascades' decision makers to drill down further on what triggered replies to their competitor's tweets and why these replies were triggered. This would then help Cascades' strategists to implement communication, PR and operational strategies that would appeal to the environmentally conscious customers and shareholders of the pulp and paper industry. Hence, giving them a competitive edge over Resolute Forest Products and other similar competitors.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.4.4 Usage of Most Common Words Over Time

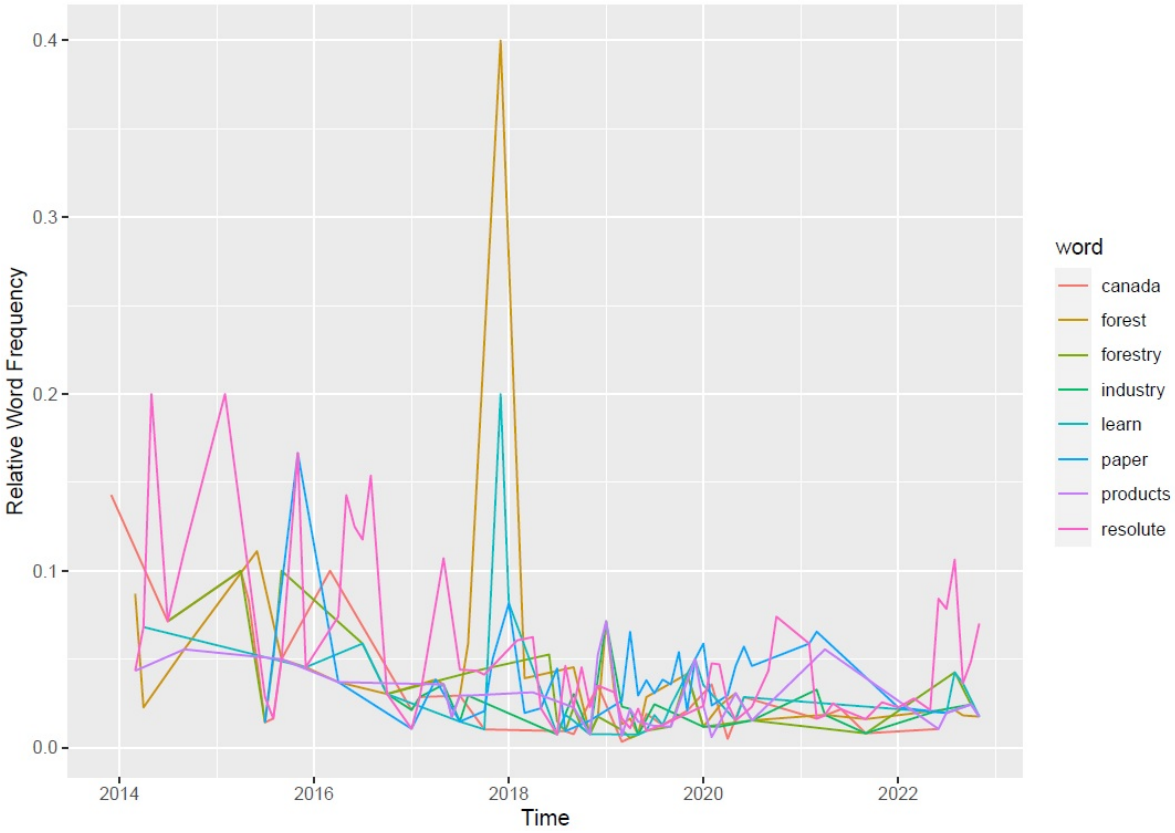


Figure 63: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts. Figure 63 is useful to highlight the change in Resolute Forest Products' focus in their public twitter messages over time.

It is clear here that the word "learn" spikes at the same time as some of the most prominent topics that Resolute Forest Products is trying to prove or disprove (i.e. learn more by clicking on the hyperlink accompanying the tweet body). One example of that is the inseparable spikes for the word "forest" as well as the word "learn" around the end of the year 2017.

This figure could be used by Cascades to track the topics that Resolute Forest Products uses to influence their customers and shareholders. Or if appropriate, Cascades could tweet about their sustainable forestry practices in the Boreals (if they have any operations there) and elsewhere as a useful counter strategy.

5.4.5 Structural Topic Modelling

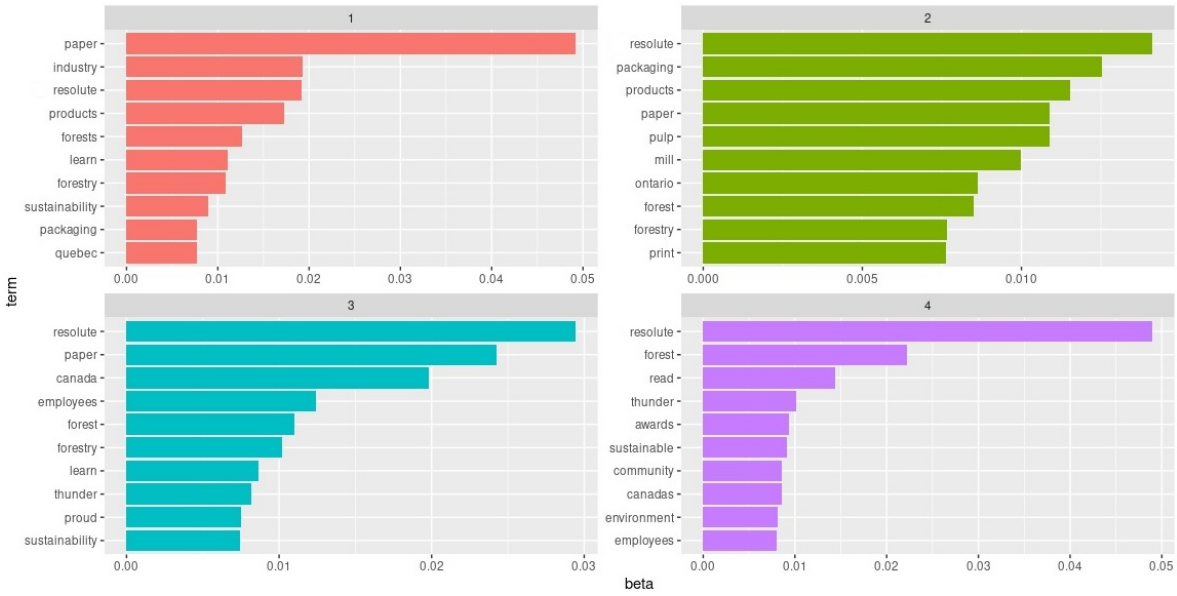


Figure 64: The optimum topic-word combination of the structural topic modelling for the tweets of Resolute Forest Product is 10 words and 4 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents Resolute Forest Products' operations and products in Quebec. As for topic 2, it presents their products and operations in Ontario. Regarding topic 3, the combination of words entail their media communications about their impact on the environment and their communities. And finally topic 4 represents their media communications about their achievements.

We would expect Resolute Forest Products to talk about their sustainable practices towards forests knowing the kind of business they are in. With that said, it may be challenging to discern when they are actually mentioning their name, Resolute Forest Products, compared to when they are referring to woodlands as forests. That is why it is always a good idea to include a relatively large number of words in the topic model to minimize these challenges.

Summing up, Resolute Forest Products' tweets are about 4 topics: Operations and products in Quebec, operations and products in Ontario, news about their impact, news about their

achievements.

5.4.6 Sentiment Analysis

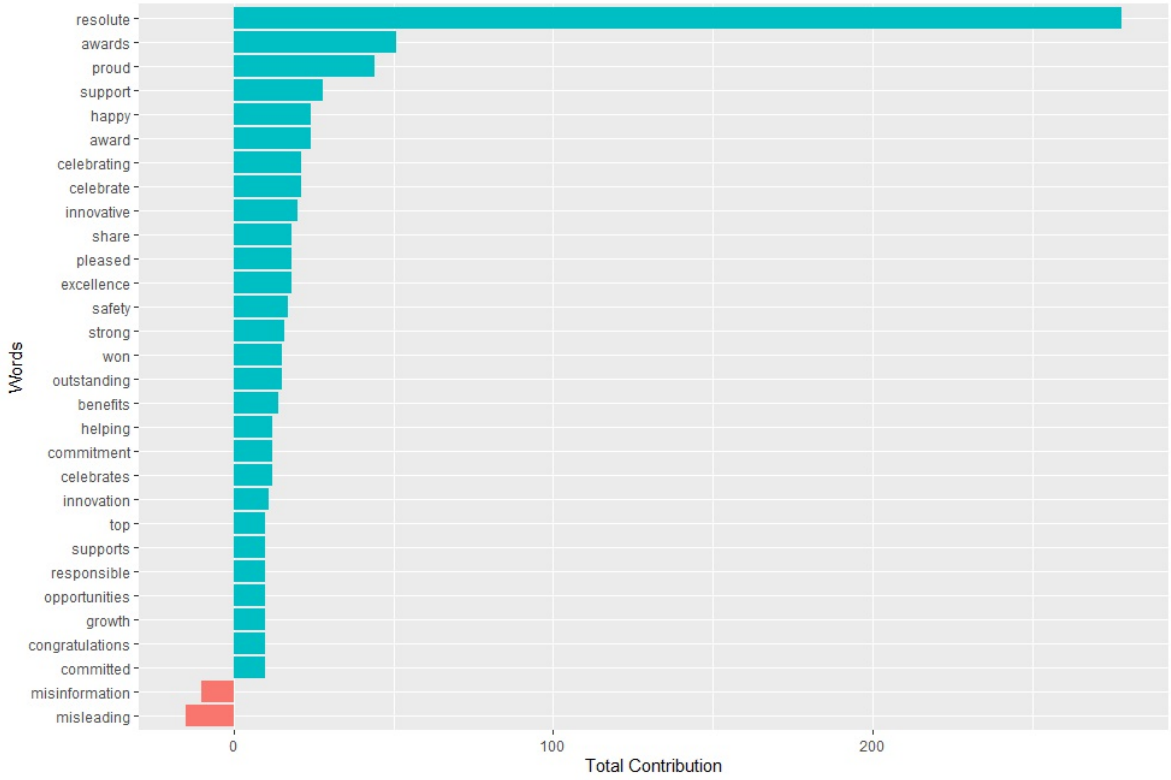


Figure 65: Sentiment analysis for Resolute Forest Products' qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Resolute Forest Products. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect Resolute Forest Products to use positive words when they publish news about the 4 topics we presented in the previous section in order to attract more customers along with more shareholders. This is clearly proven by the top 5 unique words that have the highest contribution to International Paper's overall message sentiment, namely: resolute, proud, awards, support, happy. Recall that these words are very similar to the top 5 words used by International Paper as well. We can also see that the words "misinformation" and "misleading" show up as the top negative contributors. But in reality, a drill down into their tweets containing both of these words indicates that they were used in the context of a positive message to disprove the bad publicity they had on the

topic of the Canadian Boreal forests. This provides a further proof to our discussion in the previous subsection. When correlated with the other data science powered exploration we performed, Cascades' strategy room could use all that to fortify their competitive position by showing strength where their competitor is weak.

Furthermore, we find that running a sentiment analysis on the crowds' replies to the tweets of Resolute Forest Products would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

Negation Used	**Word Used**	**Score**	**Number of occurrences**	**Comments**
dont	have	-3	1	Synonym of lack
no	wonder	N/A	1	Suspect sarcastic connotation
no	charge	+2	1	Synonym of free
no	substitute	-3	1	Synonym of unique
not	mean	-3	2	Could be taken positively | Used to counter argument or elaborate
not	mutually	N/A	1	Could be taken positively |
not	trees	-5	1	Not environmentally friendly
not	believe	-5	1	Accusation of relaying untrue information
not	canada	N/A	1	Cannot rank a country Suspect context talks about specifying location
dont	let	-2	1	Could be taken positively |
not	sustainable	-5	1	Synonym of undurable
Total	**Score**	**-27**		

A sentiment analysis figure was not produced for the negated words even though there exists

12 instances of negation for 11 tokens, namely: have, wonder, charge, substitute, mean (negated twice), mutually, believe, trees, Canada, let and sustainable. The reason behind that is because these words do not exist in the afinn lexicon we used for our sentiment analysis. Therefore, the algorithm would produce a contribution of zero for each of these tokens resulting in no diagram at all. Regardless, these words could be manually added to the afinn lexicon with an objectively estimated score for each of them to follow the afinn lexcion methodology. But if we were to be the devil's advocate and take the usable words at their worst meaning possible, it would not affect our overall score for Resolute Forest Product's tweets that we can see in the previous diagram. The table on the previous page presents how we would rate the words in addition to our comments.

5.4.7 Overall Network of Words in Tweets

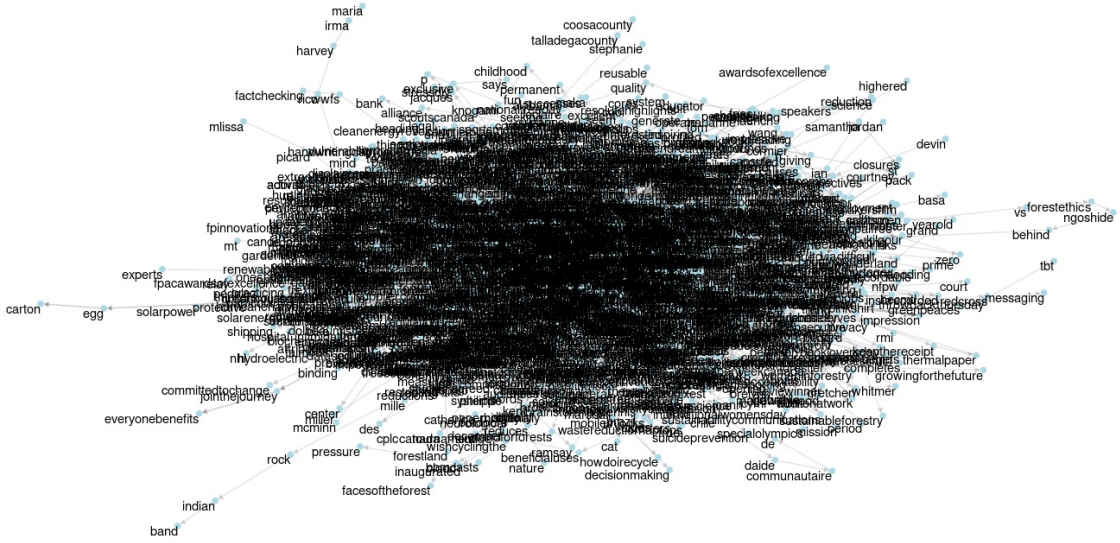


Figure 66: This directed network graph is showing the overall communication strategy of Resolute Forest Products on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Resolute Forest Products structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of Resolute Forest Products. This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd. This also helps Cascades to fill up any structural holes in their communication and to uncover what Resolute Forest Products talk

about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.4.8 Network of Most Common Used Pairs of Words

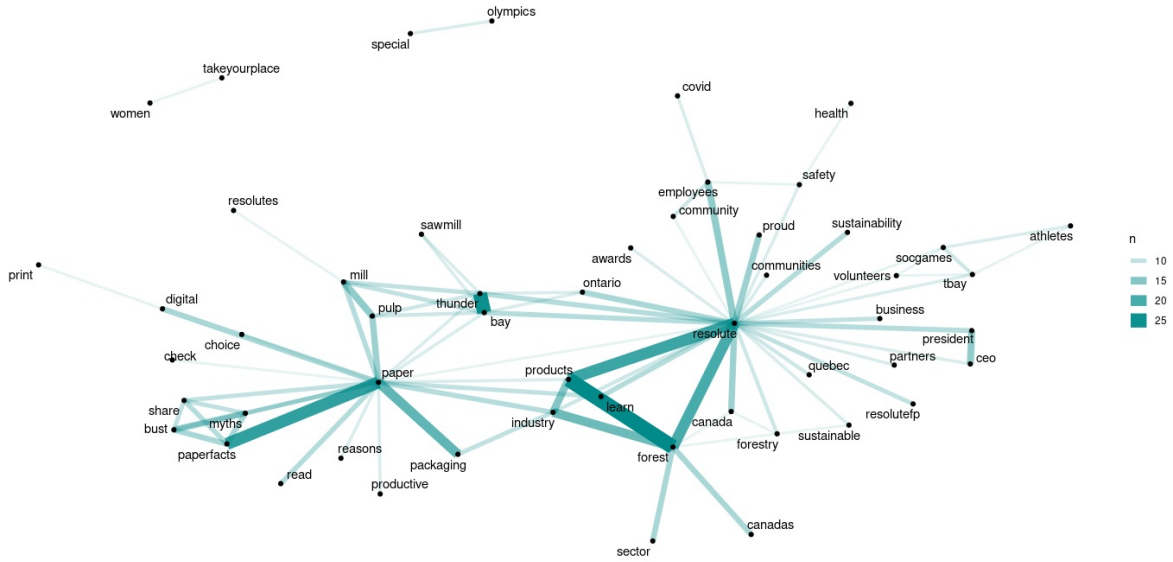


Figure 67: This network graph is showing the most common usage of the pair of words tweeted by Resolute Forest Products. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that Resolute Forest Products tweets about. We can clearly see the presence of 2 main quasi star shaped graphs. The quasi star on the right has an operational theme while the star on the left has more of a product theme. We can also see that Thunder Bay in Ontario houses the majority of the operations of Resolute Forest Products. Another point to highlight is the complete network on the far left that shows how Resolute Forest Products tries to counter what they refer to as "misinformation" by creating "paperfacts".

It would be useful for Cascades to drill down further into Resolute Forest Products' operations as well as their presented products in order to identify opportunities that will help them gain a competitive advantage.

5.5 Sappi Group

Sappi Group (`@sappigroup`) present themselves on twitter as a leading global provider of sustainable woodfiber products and innovative solutions (#sustainability #sustainablesolutions #innovation). They are based out of Johannesburg (South Africa) and joined Twitter in November 2017. The tweets we were able to collect for `@sappigroup` span dates from November 2017 till November 2022.

5.5.1 Exploratory Data Analysis

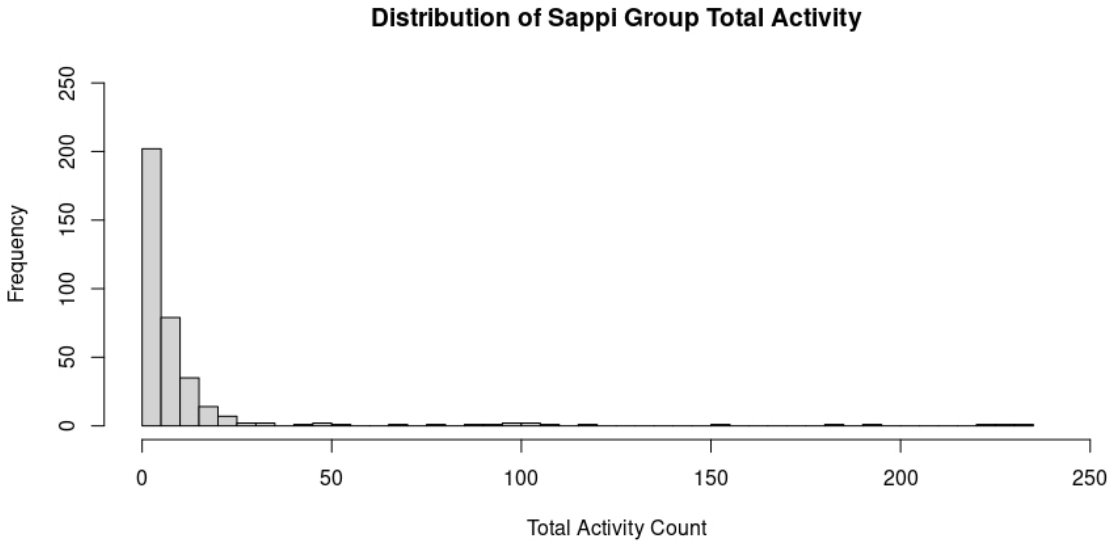


Figure 68: Distribution of total activity counts for the different tweets by Sappi Group

Looking at the Twitter activity distribution of Sappi Group, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by Sappi Group that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 10 for Sappi Group, we will take all the tweets whose total activity is larger than 10. This brings down the total number of tweets by Sappi Group from 361 to 80 tweets for analysis.

5.5.2 Identification and Classification of Spikes

Figure 69: Graphs showing spikes in total activity for Sappi Group's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Sappi Group did a better job in keeping the twitter crowd engaged in the first half of the period of analysis compared to the second half. But according to the total activity count graph, the twitter crowd found that Sappi Group tweeted about hotter topics during the second half of the period of analysis. Hence, the presence of more significant spikes starting from the year 2020 onward. Correlating all 4 graphs together, all the significantly "spiky" tweets attracted different combinations of replies, retweets and likes with only one tweet strictly attracting likes towards the end of the year 2022. In addition, the first ternary ratio graph shows a low concentration of replies through the period of analysis. This overview of engagements tells us that the tweets made by Sappi Group are generally endorsed by the twitter crowd and are viewed as credible.

In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down. The next section will help us understand which words attracted each type of reaction over the whole period of analysis.

5.5.3 Frequency of Most Common Words per Reaction

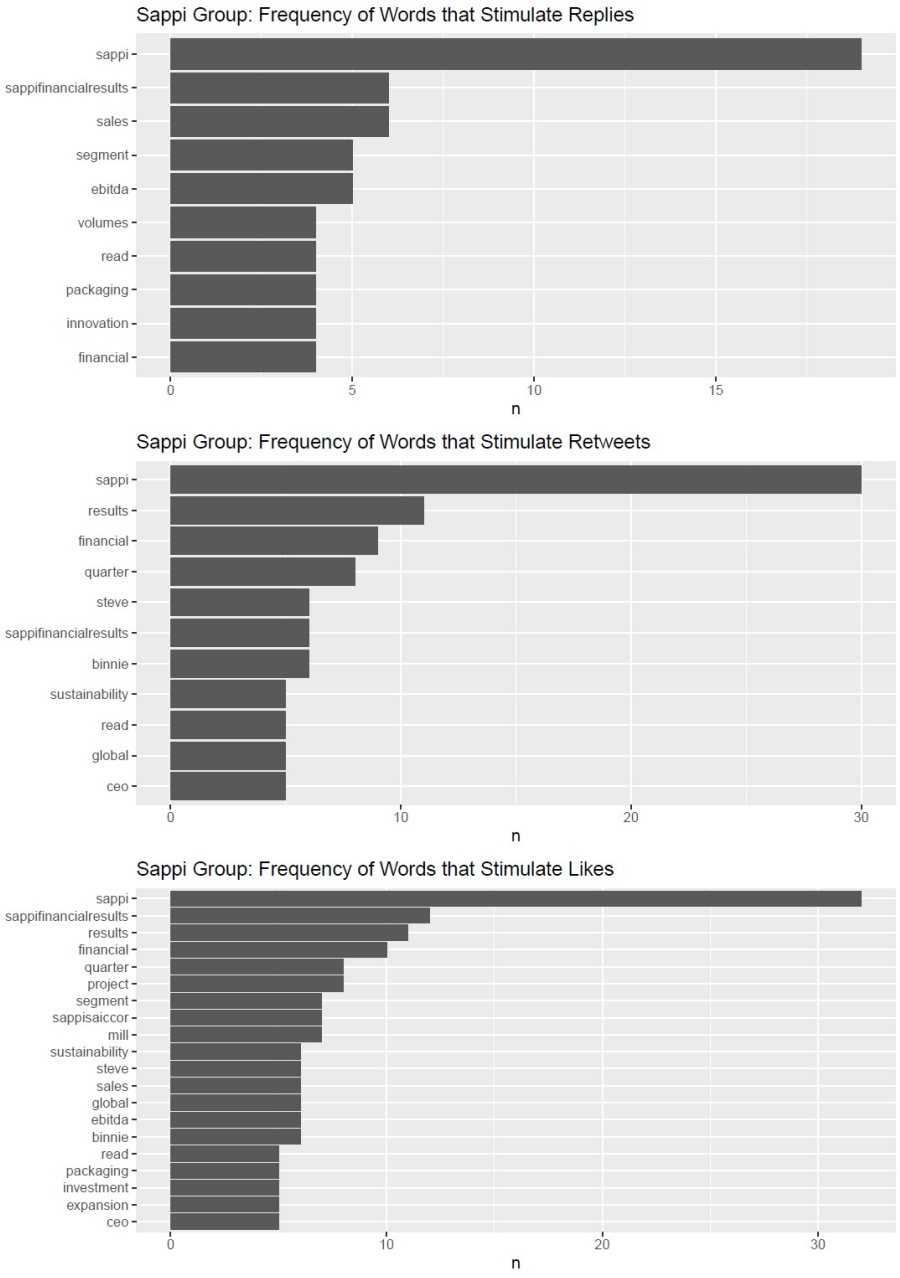


Figure 70: Frequency of words in the tweets of Sappi Group on hot topics

As shown in the figure above, all 3 bar charts have the financial results of Sappi Group as the main theme using the following most frequent words: sappi, financial, results, ebitda and quarter. While innovation showed up as part of the most common words that attracted replies, sustainability showed up as part of the most common words that attracted retweets and likes. Moreover, we can see that the most common words that attracted likes include those that attracted retweets with further additions that discuss Sappi Group's investments and expansion plans, even though these 2 words appeared relatively much less.

Learning what triggers replies VS. retweets VS. likes to the tweets of Sappi Group could help the decision makers of Cascades to tailor their strategies by learning more about the different topics discussed by Sappi Group and how these topics are received by the different stakeholders.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.5.4 Usage of Most Common Words Over Time

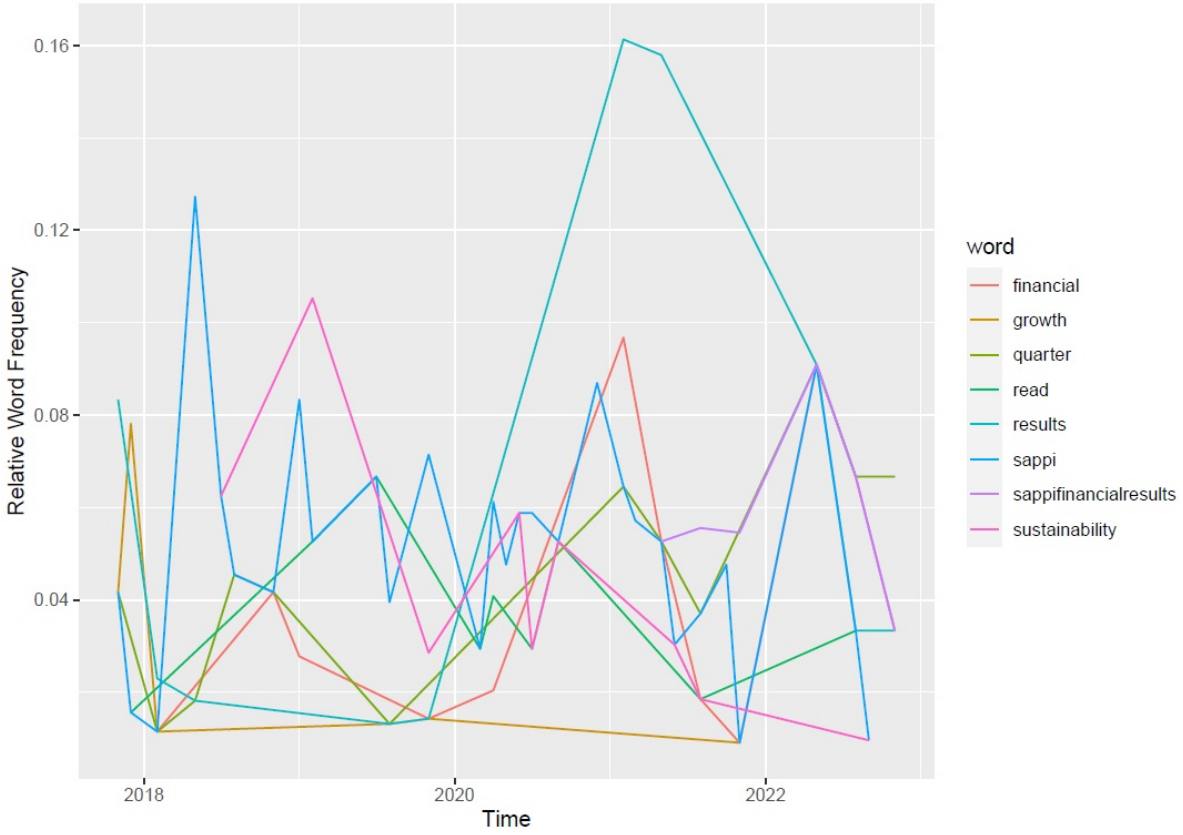


Figure 71: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this

measure would help us perform equitable analysis for all the different twitter accounts. Figure 71 is useful to highlight the change in Sappi Group's focus in their public twitter messages over time.

We can see that all of the most common words shown in the graph refer to the financial results of Sappi Group except for the topic of sustainability. The word sustainability was more dominant during the first half of the period of analysis whereas the financial results were more dominant during the second half of the period of analysis. We can also see that Sappi Group shared intermittently with a low relative frequency about their growth plans. But they were more bold in tweeting about their financial results throughout the whole period, especially around the beginning of 2021. It would be interesting to find out the reason behind this large focus on financial results in the next subsections.

This figure could be used by Cascades to track the topics that Sappi Group uses to influence their customers and shareholders and to identify the upcoming trends in their competitor's behavior. They can also use this graph to implement strategies and communication that overshadow Sappi Group's tweets.

5.5.5 Structural Topic Modelling

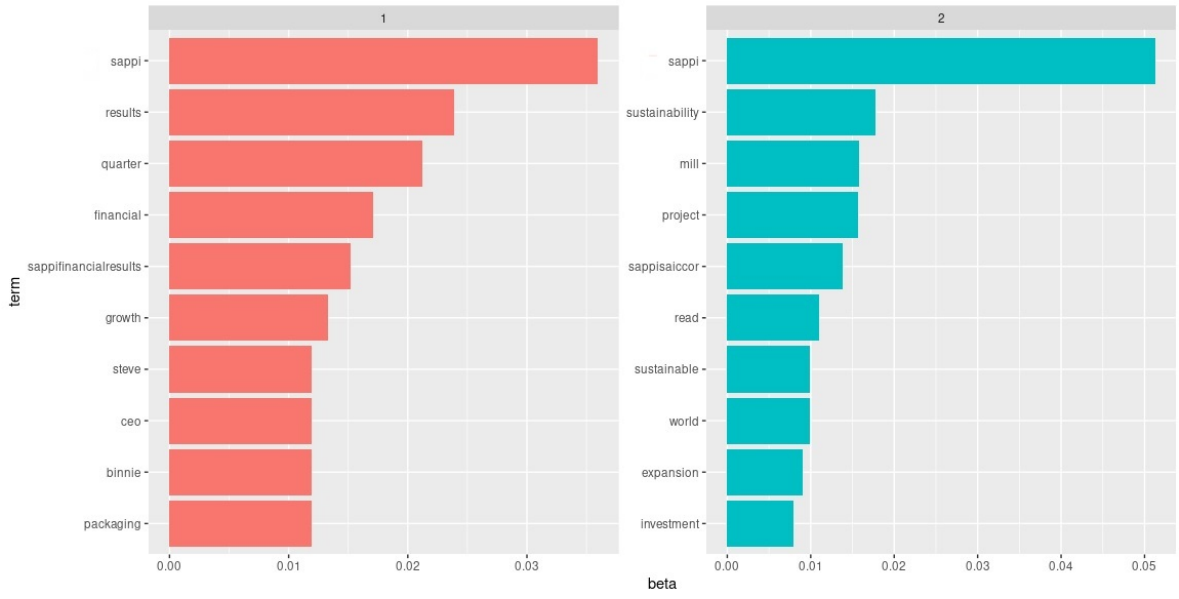


Figure 72: The optimum topic-word combination of the structural topic modelling for the tweets of Sappi Group is 10 words and 2 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents news about Sappi Group's financial results to attract shareholders. As for topic 2,

it represents news about Sappi Group's operations and practices in their business.

Summing up, Sappi Group's tweets are about 2 topics: Financial results and operations.

5.5.6 Sentiment Analysis

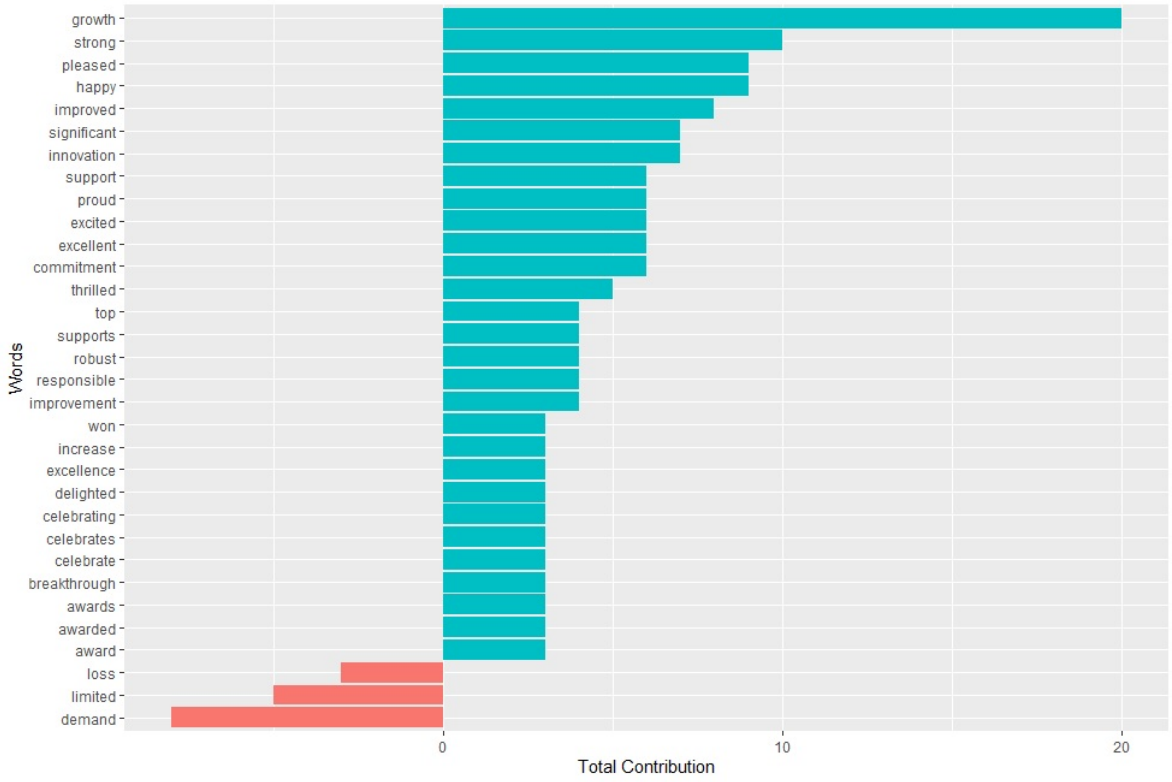


Figure 73: Sentiment analysis for Sappi Group's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Sappi Group. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect Sappi Group to use positive words when they publish news about the 2 topics we presented in the previous section in order to attract more customers along with more shareholders. It is important to highlight that the most frequent words in Sappi Group's tweets discuss their financials results which are more appealing to shareholders than customers. This is clearly proven by the top 5 unique words that have the highest contribution to Sappi Group's overall message sentiment, namely: growth, strong, happy, pleased and improved. We can also see that the words loss, limited and demand show up as the top negative contributors. But in reality, a drill down into their tweets containing

these words shows that Sappi Group's team turned loss in 2020 to profit in 2021. This discovery, coupled with the results from the previous subsections, explains why Sappi Group was tweeting heavily about their financial results and why we had a spike in the relative frequency of the word "results" at the beginning of 2021. As for the word "limited", it was used to indicate the limited personal liability of the company's shareholders, whereas "demand" was used in an economic context. Hence, a positive connotation after all.

Furthermore, we find that running a sentiment analysis on the crowds' replies to the tweets of Sappi Group would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

Negation Used	**Word Used**	**Score**	**Number of occurrences**	**Comments**
no	other	+3	1	Synonym to unique
no	shortterm	-5	1	No short-term solution
no	further	-4	1	Synonym to closer Could be used positively
Total	**Score**	**-6**		

A sentiment analysis figure was not produced for the negated words even though there exists 3 instances of negation for 3 tokens, namely: other, shortterm, further. The reason behind that is because these words do not exist in the afinn lexicon we used for our sentiment analysis. Therefore, the algorithm would produce a contribution of zero for each of these tokens resulting no in diagram at all. Regardless, these words could be manually added to the afinn lexicon with an objectively estimated score for each of them to follow the afinn lexcion methodology. But if we were to be the devil's advocate and take the usable words at their worst meaning possible, their effect would be negligible to our overall score for Sappi Group's tweets that we can see in the previous diagram. The table above is a quick summary of how we would rate the words in addition to our comments.

5.5.7 Overall Network of Words in Tweets

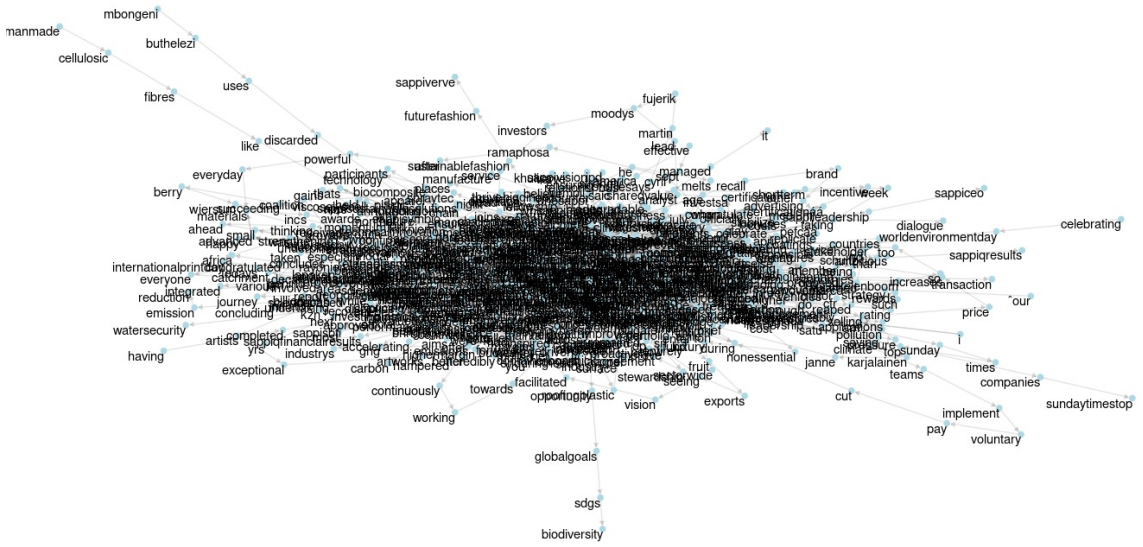


Figure 74: This directed network graph is showing the overall communication strategy of Sappi Group on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Sappi Group structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of Sappi Group. This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd. This also helps Cascades to fill up any structural holes in their communication and to uncover what Sappi Group talks about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.5.8 Network of Most Common Used Pairs of Words

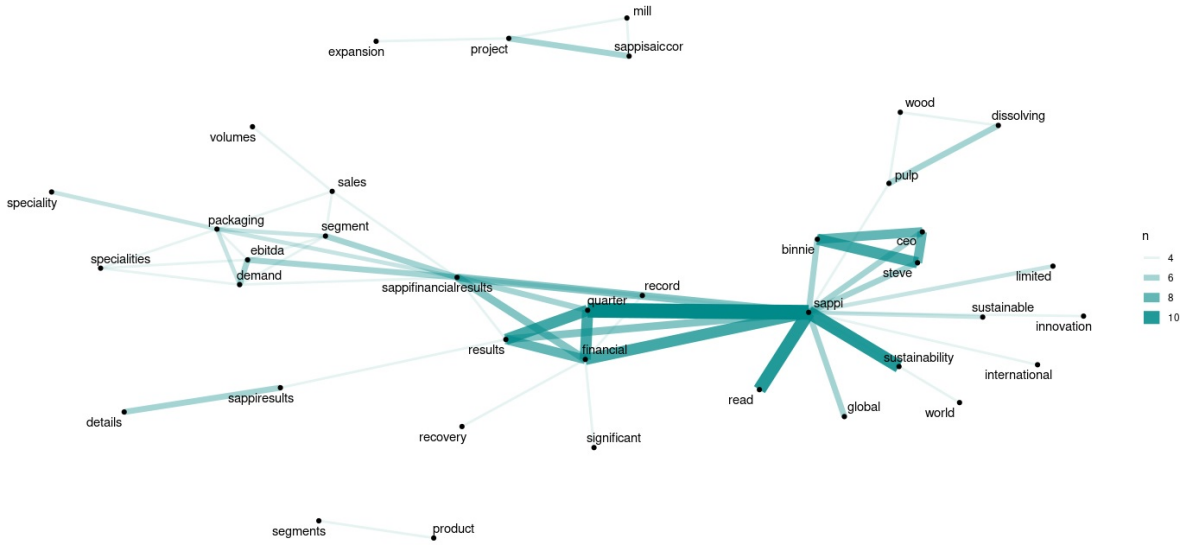


Figure 75: This network graph is showing the most common usage of the pair of words tweeted by Sappi Group. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that Sappi Group tweets about. We can clearly see the presence of 2 main quasi star shaped graphs. From the thickness and darkness of the edges, we can deduce that Sappi Group focused a lot on their financial results in their tweets. The left side also shows us some of the operations that contributed to Sappi Group's financial results. We can also see that Sappi Group focuses more on the sustainability of their practices compared to the innovation aspect. It is interesting to note how the algorithm generated graph recognized that the CEO of Sappi Limited is Steve Binnie.

It would be useful for Cascades to drill down further into Sappi Group's operations as well as their presented products in order to identify opportunities that will help them gain a competitive advantage.

5.6 Sappi Southern Africa

Sappi Southern Africa (`@sappisoutherna`) present themselves on twitter as a leader in sustainable woodfiber products and solutions. They are based out of South Africa and joined Twitter in December 2017. It is important to note that they fall under the umbrella of Sappi. The tweets we were able to collect for `@sappisoutherna` span dates from December 2017 till November 2022.

5.6.1 Exploratory Data Analysis

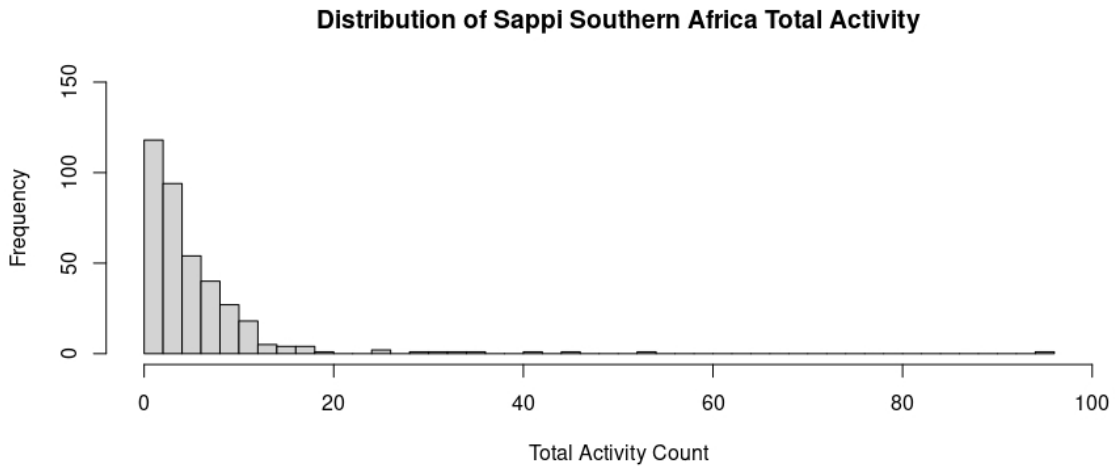


Figure 76: Distribution of total activity counts for the different tweets by Sappi Southern Africa

Looking at the Twitter activity distribution of Sappi Southern Africa, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by Sappi Southern Africa that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 7 for Sappi Southern Africa, we will take all the tweets whose total activity is larger than 7. This brings down the total number of tweets by Sappi Southern Africa from 375 to 86 tweets for analysis.

5.6.2 Identification and Classification of Spikes

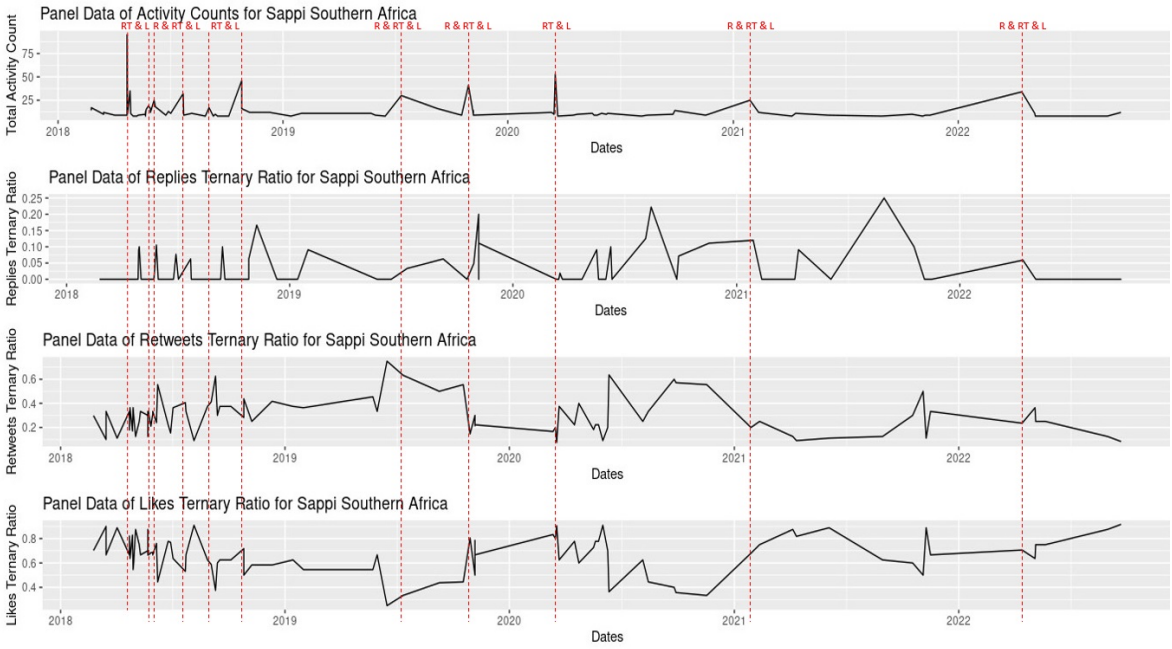


Figure 77: Graphs showing spikes in total activity for Sappi Southern Africa's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Sappi Southern Africa presented only a few hot topics to the twitter crowd per the spikes in total activity count over the whole period of analysis. It is also seen that there are sudden spikes such as the one shown during the first quarter of the year 2020, as well as gradual spikes during the first quarter of 2021 or 2022. Sudden spikes indicate a topic that intrigued the interest of the twitter crowd for a very short period of time. Whereas a gradual spike may either indicate the lack of enough data points to draw a more detailed smooth curve, or a conversation that started building up through several tweets over time before it reached a peak. Indeed, a closer look at the tweet with a sudden spike during the first quarter of 2020, Sappi Southern Africa tweeted about #GlobalRecyclingDay and how they have always contributed to the circular economy to celebrate that one day. Thus leading to 1 reply, 4 retweets and 48 likes from the twitter crowd reacting to this tweet. Whereas the gradual spike tweets in this case were due to the lack of enough distinct data points after our elimination of all the tweets in the previous subsection that were perceived as "unimportant" by the twitter crowd. Correlating all 4 graphs together, all the significantly "spiky" tweets attracted different combinations of replies, retweets and likes with no tweets attracting just a single type of reaction. In addition, the scale of the first ternary ratio graph shows a very low

concentration of replies throughout the period of analysis. This overview of engagements tells us that the tweets made by Sappi Southern Africa are generally endorsed by the twitter crowd and are viewed as credible.

In Cascades' strategy room, this graph could be used to guide the decision makers on the time intervals that require a deeper drill down. The next section will help us understand which words attracted each type of reaction over the whole period of analysis.

5.6.3 Frequency of Most Common Words per Reaction

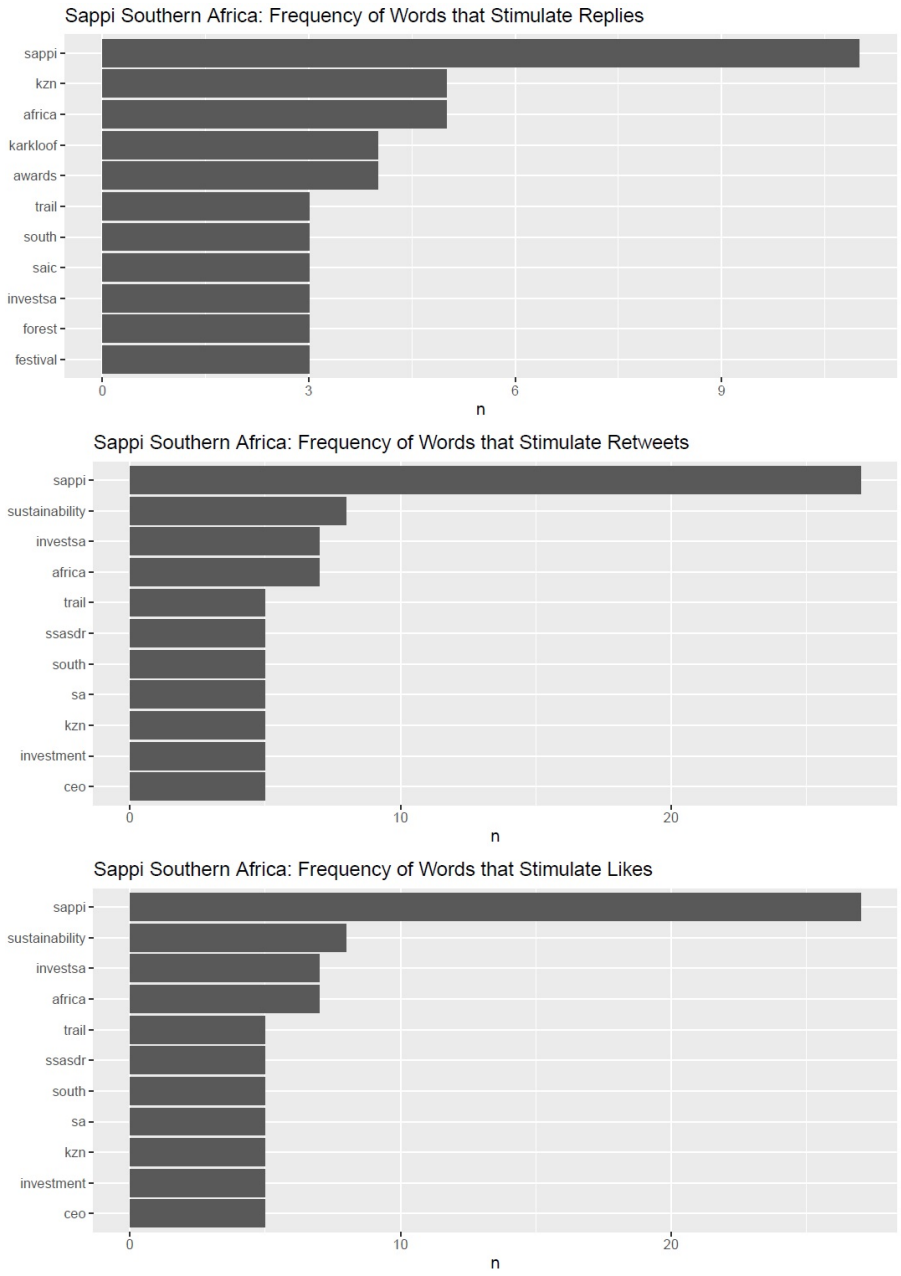


Figure 78: Frequency of words in the tweets of Sappi Southern Africa on hot topics

As shown in figure 78 on the previous page, all 3 bar charts include words related to the investment strategy of Sappi Southern Africa. While kzn (stands for KwaZulu Natal which is a province in South Africa), karkloof (refers to the forest part of a Natural reserve in South Africa) and awards dominated Sappi Southern Africa's tweets that attracted replies, sustainability dominated their tweets that attracted retweets and likes. This is a clear indication that Sappi Southern Africa's sustainability efforts are endorsed by the twitter crowd and viewed as credible. Combining all these pieces of information could indicate to Cascades the difficulty level of penetrating the South African market due to Sappi Southern Africa's heavy investments there and their ability to raise the liability of foreignness along with the CAGE distances (Pankaj. Ghemawat 2007), notably the cultural and administrative aspects. but this requires a further drill down to confirm.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.6.4 Usage of Most Common Words Over Time

Figure 79 on the previous page shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous section. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts. This figure is useful to highlight the change in Sappi Southern Africa's focus in their public twitter messages over time.

We can see that besides the name and operational location of Sappi Southern Africa, the theme of sustainability was always present throughout the whole period of analysis. This could be used by Cascades to track how the topics tweeted by Sappi Southern Africa evolved over time to influence their customers and shareholders in addition to identifying the upcoming trends in their competitor's behavior.

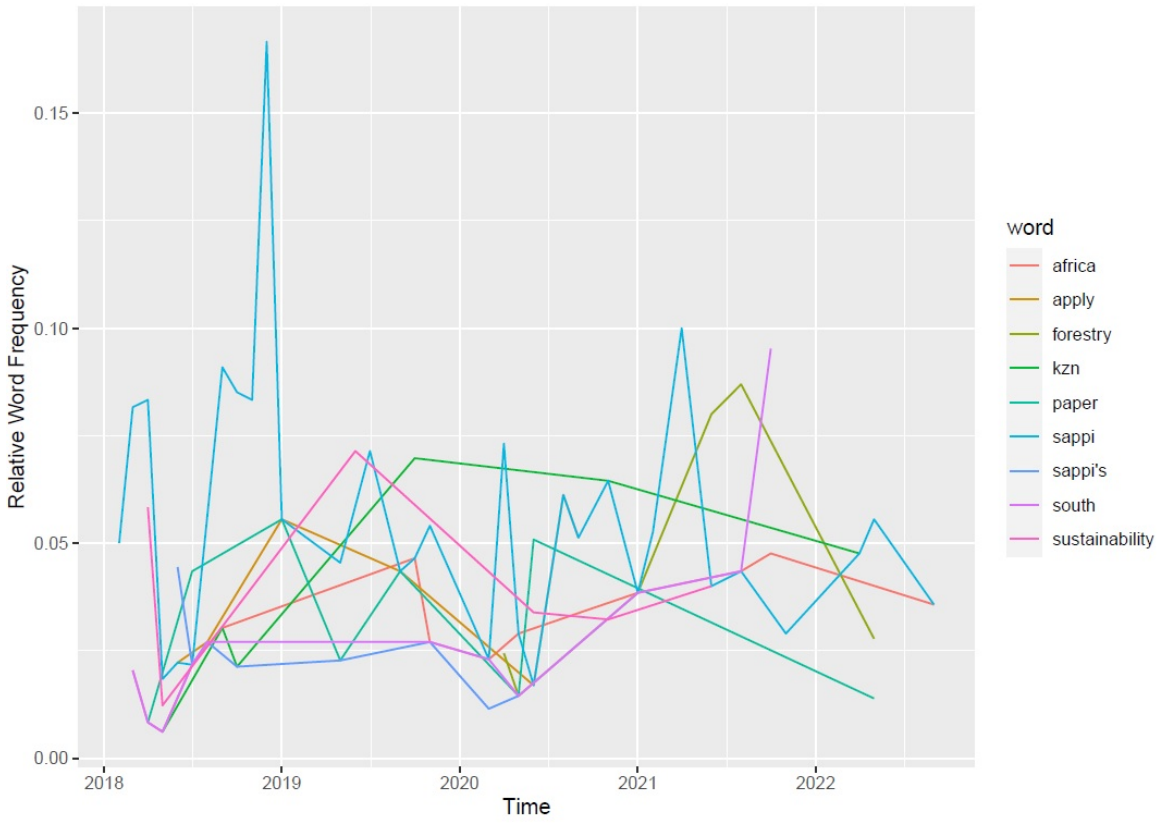


Figure 79: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

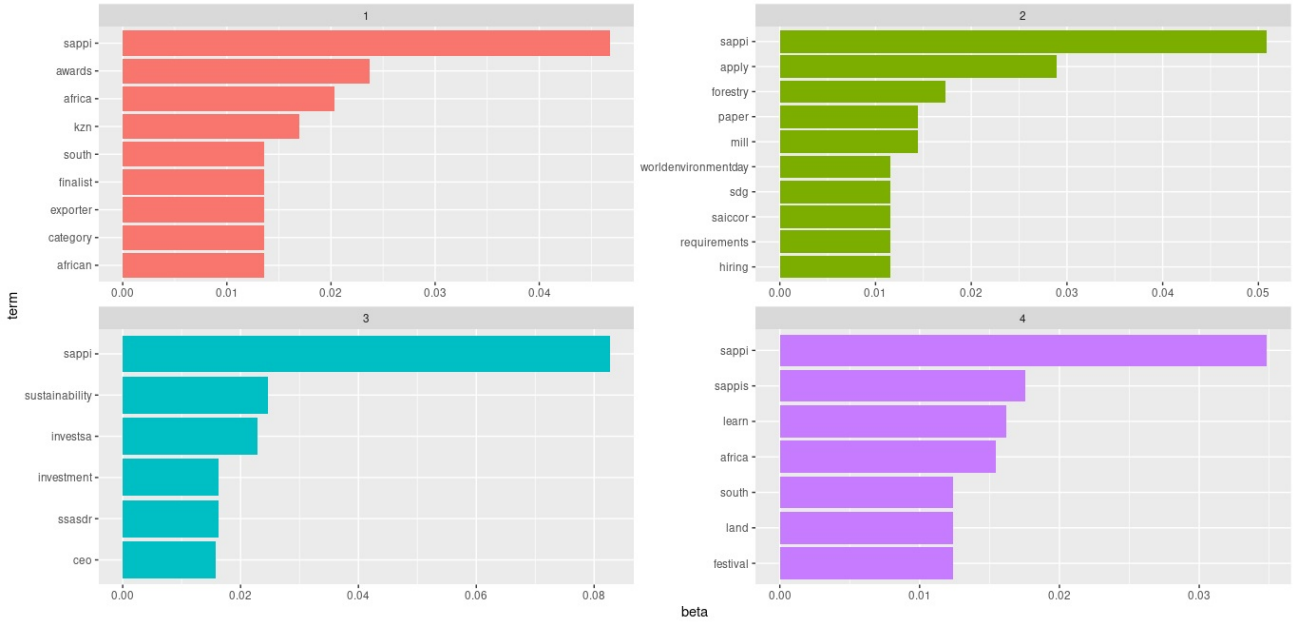


Figure 80: The optimum topic-word combination of the structural topic modelling for Sappi South Africa's tweets is 10 words and 4 topics

5.6.5 Structural Topic Modelling

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents Sappi Southern Africa's tweets about their awards and achievements. As for topic 2, it presents how Sappi Southern Africa attracts talent by tweeting about their company values with their invitation for candidates to apply. Regarding topic 3, it represents how they talk about their investments in South Africa. Finally, topic 4 talks encompasses tweets that publish news on their activities and operations.

Summing up, Sappi Southern Africa tweets about 4 topics: Their awards and achievements, hiring, their investments in South Africa along with their activities and operations.

5.6.6 Sentiment Analysis

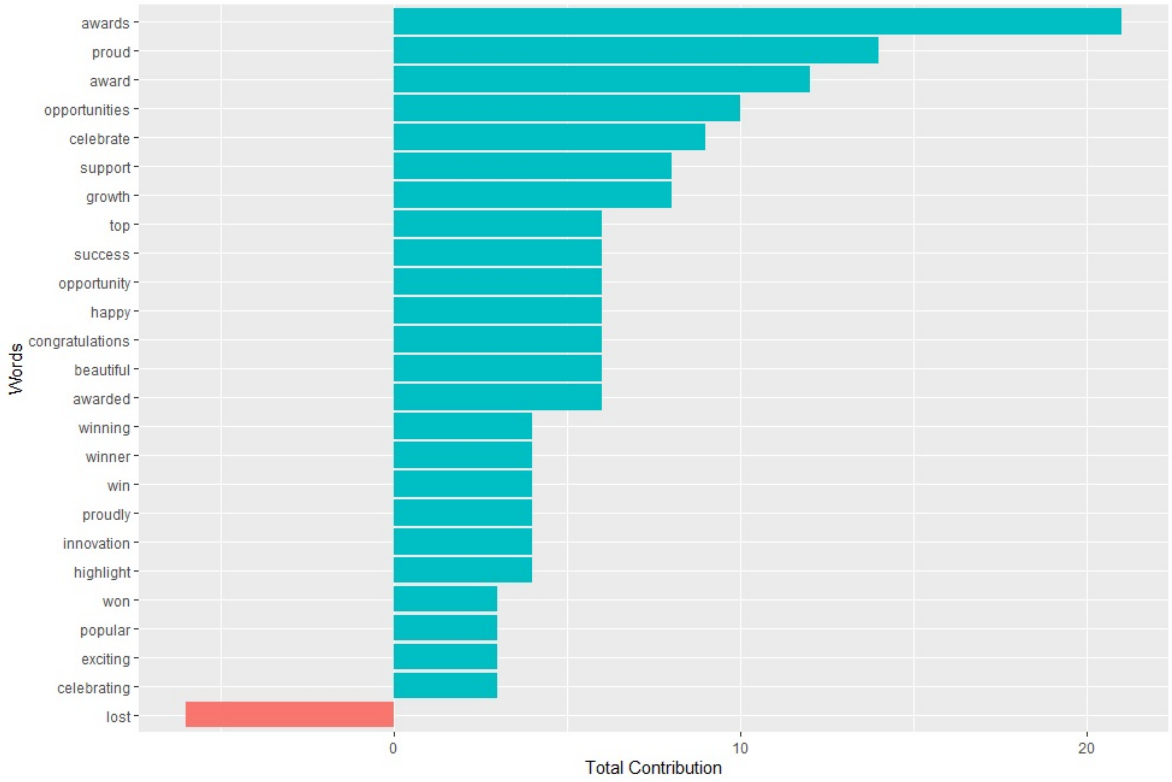


Figure 81: Sentiment analysis for Sappi Southern Africa's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Sappi Southern Africa. As a for-profit pulp and paper company that cares about their public image

within the competitive landscape, we would expect Sappi Southern Africa to use positive words when they publish news about the 4 topics we presented in the previous subsection in order to attract more customers along with more shareholders. It is important to highlight that the most frequent words in Sappi Southern Africa's tweets discuss their investments in South Africa and in their sustainable practices which are equally appealing to both shareholders and customers.

As expected, the top 5 unique words that have the highest contribution to Sappi Southern Africa's overall message sentiment boast about their pride and achievements, namely: award, proud, opportunities, celebrate and support. We can also see that the word lost shows up as the top negative contributor. A deeper drill down into the tweets containing that negative word shows that Sappi Southern Africa was holding events to commemorate the team members they lost due to COVID19. Hence, a true negative connotation after all.

Furthermore, we find that running a sentiment analysis on the crowds' replies to the tweets of Sappi Southern Africa would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

Negation Used	**Word Used**	**Score**	**Number of occurrences**	**Comments**
without	it	-3	1	Synonym to alone Could be positive
not	just	-5	1	Refers to additions Could be positive
Total	**Score**	**-8**		

A sentiment analysis figure was not produced for the negated words even though there exists 2 instances of negation for 2 tokens, namely: it and just. The reason behind that is because these words do not exist in the afinn lexicon we used for our sentiment analysis. Therefore, the algorithm would produce a contribution of zero for each of these tokens resulting no in diagram at all. Regardless, these words could be manually added to the afinn lexicon with an objectively estimated score for each of them to follow the afinn lexcion methodology. But if

we were to be the devil's advocate and take the usable words at their worst meaning possible, their effect would be negligible to our overall score for Sappi Southern Africa's tweets that we can see in the previous diagram. The quick table above shows how we would rate the words with our comments included.

5.6.7 Overall Network of Words in Tweets

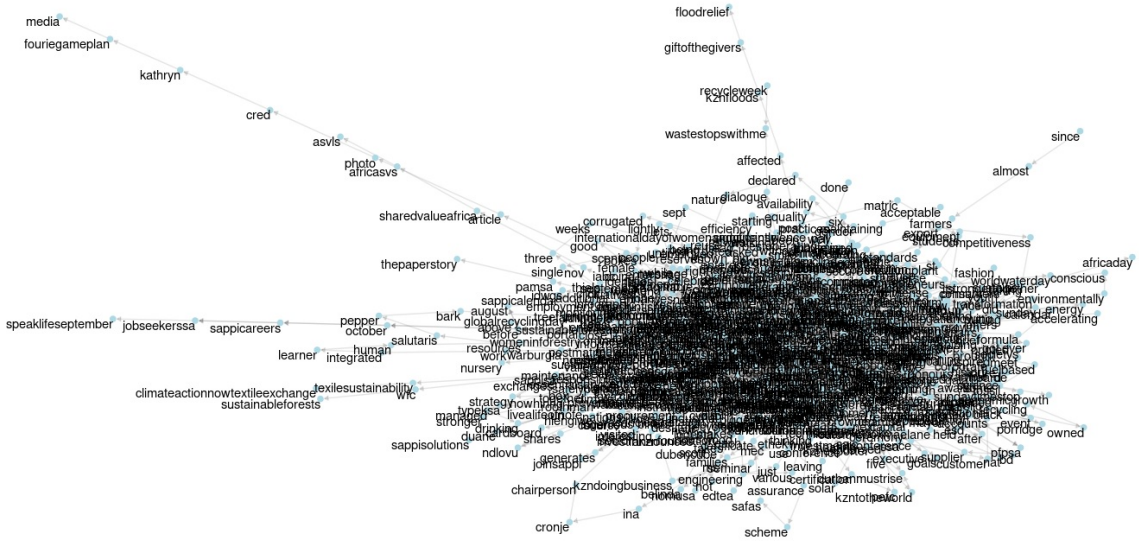


Figure 82: This directed network graph is showing the overall communication strategy of Sappi Southern Africa on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Sappi Southern Africa structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of Sappi Southern Africa. It is interesting to see how different topics start far away from the dense part of the network until they converge into that dense center. A good example of that is Sappi Southern Africa's sustainable practices that start to converge from the top of the dense network. Another example of that is their focused PR efforts that start to converge from the top left of the dense network. An additional example is their talent attraction effort that starts converging from the left side of the dense network. The convergence of all the different topics to one of few clusters of words within the dense area of the network shows how all of Sappi Southern Africa's practices tie back to their values or mission.

This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd while also converging to their values and mission. This also helps Cascades to fill up any structural holes in their communication and to uncover what Sappi Southern Africa talks about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.6.8 Network of Most Common Used Pairs of Words

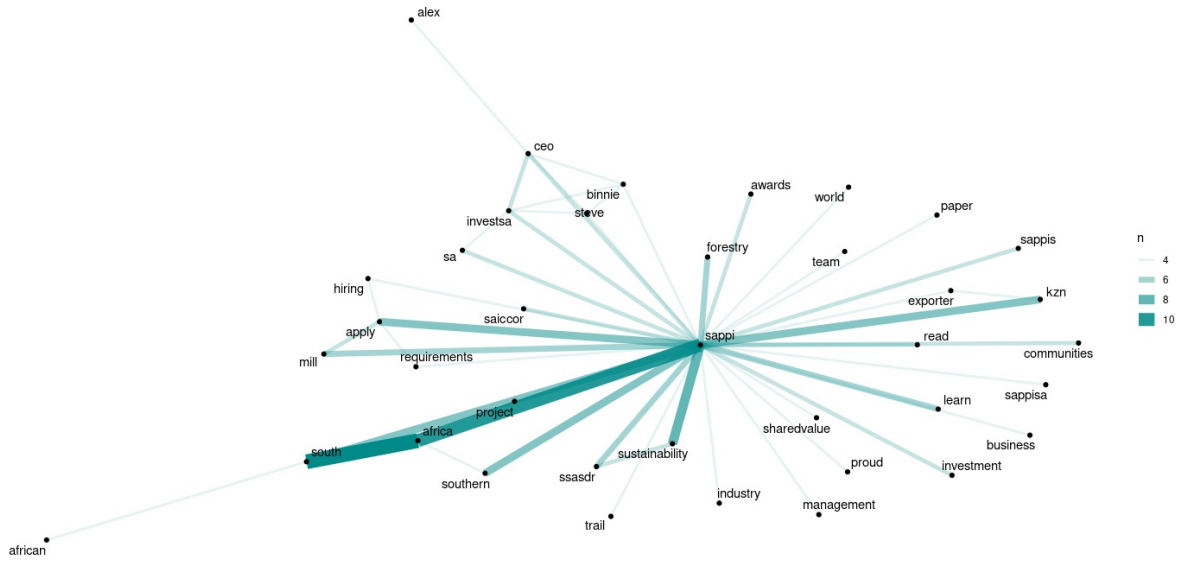


Figure 83: This network graph is showing the most common usage of the pair of words tweeted by Sappi Southern Africa. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that Sappi Southern Africa tweets about. We can clearly see the presence of 1 main quasi star shaped graph. It is interesting to see how this graph further confirms our deduction from figure **??** in the previous section about how Sappi Southern Africa's communication strategy always ties everything back to Sappi, the name that represents their values and mission. It is also interesting to note how the algorithm generated graph recognized that the CEO of Sappi Limited is Steve Binnie.

It would be useful for Cascades to drill down further into Sappi Southern Africa's operations in order to identify opportunities that will help them gain a competitive advantage.

5.7 Stora Enso

Stora Enso (`@storaenso`) present themselves on twitter as part of the bioeconomy, a leading global provider of renewable solutions in packaging, biomaterials, wooden construction and paper. They have not specified where they are based out of but their headquarters are in Helsinki (Finland) and they joined Twitter in August 2009. The tweets we were able to collect for `@storaenso` span dates from April 2011 till November 2022.

5.7.1 Exploratory Data Analysis

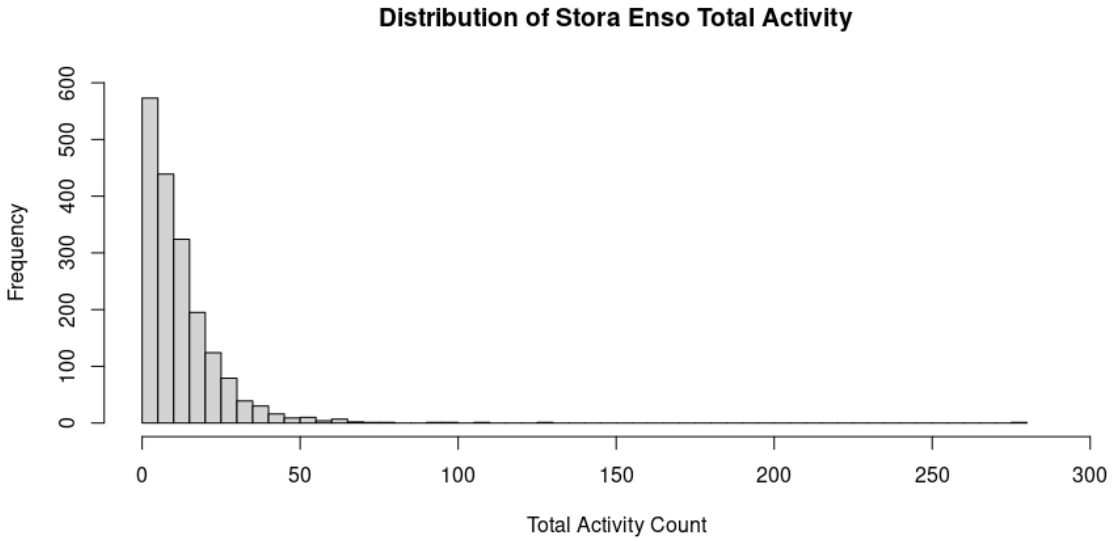


Figure 84: Distribution of total activity counts for the different tweets by Stora Enso

Looking at the Twitter activity distribution of Stora Enso, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by Stora Enso that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 17 for Stora Enso, we will take all the tweets whose total activity is larger than 17. This brings down the total number of tweets by Stora Enso from 1,858 to 428 tweets for analysis.

5.7.2 Identification and Classification of Spikes

Figure 85: Graphs showing spikes in total activity for Stora Enso's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Stora Enso presented only a few hot topics to the twitter crowd per the spikes in total activity count over the whole period of analysis. Another point to mention is how Stora Enso started doing a relatively good job in keeping the twitter crowd engaged from the beginning of the year 2017. And according to the total activity count graph, the twitter crowd found that Stora Enso tweeted about hotter topics during the second half of the period of analysis compared to the first half leading to the significant spikes seen. We can also see that retweets were always accompanied by likes which tells us that the words attracting retweets will be similar, if not identical, to the words in the tweets that attracted likes. Correlating all 4 graphs together, all the significantly "spiky" tweets attracted different combinations of replies, retweets and likes with no tweets attracting just a single type of reaction. In addition, the low scale of the first ternary ratio graph shows a very low concentration of replies throughout the period of analysis. This overview of engagements tells us that the tweets made by Stora Enso are endorsed by the twitter crowd and are viewed as credible.

Following up on that last note and as can be seen around the first quarter of 2022, a very significant spike in the total activity count is present and is accompanied by a very high spike

in ternary ratio of likes as well as a low spike in the ternary ratio of replies. This could be translated into Stora Enso tweeting about a topic that is considered as very hot by the twitter crowd, who also strongly agrees with the contents of their tweet. Indeed, a deeper drill down into this particular tweet shows that Stora Enso announced they will be stopping all production, sales, import and export to and from Russia to show solidarity with the people of Ukraine. This tweet garnered a total of 5 replies, 39 retweets and 233 likes.

This graph and analysis is a very good example of how Cascades' strategy room could benefit from the exploratory process using data science by identifying time intervals that deserve a drill down. This graph also provides a strong proof to the points mentioned in Joshua Minot's paper that replies indicate a topic that the crowd disagrees with or sees as controversial while retweets and likes indicate an endorsed topic (Minot 2022). What follows will help us understand which words attracted each type of reaction over the whole period of analysis.

5.7.3 Frequency of Most Common Words Over Time

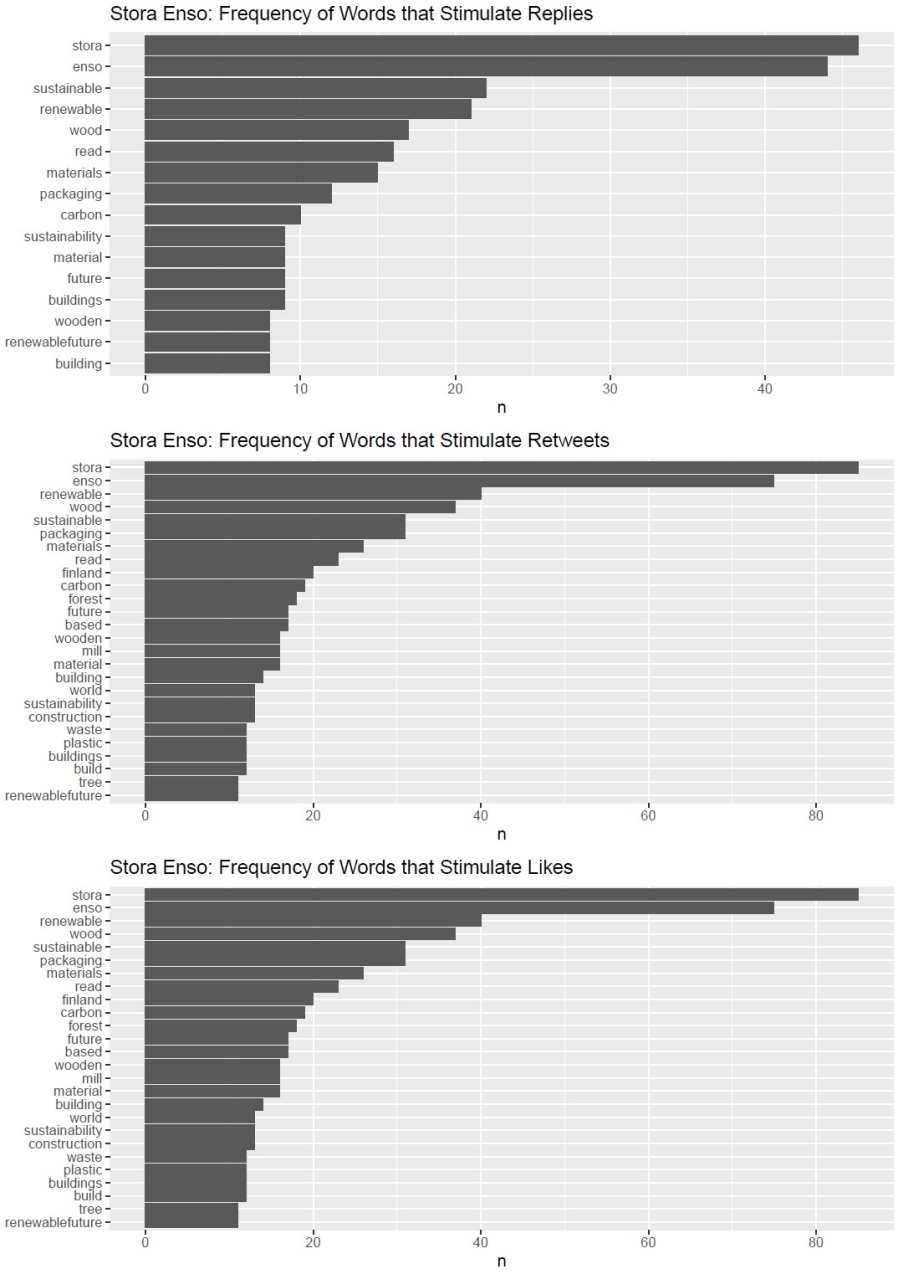


Figure 86: Frequency of words in the tweets of Stora Enso on hot topics

As shown in the figure above, the words stora, enso, sustainable, renewable and wood were among the top 5 words that attracted all 3 types of reactions having the highest frequencies of all. Based on our analysis from the previous subsection, Stora Enso's words that attracted retweets are identical to those that attracted likes. But those words also include the words that attracted replies in addition to other words such as forest, Finland, mill, world, construction and plastic. This generally indicates that the tweets attracting replies are not controversial by definition since they attracted retweets and likes as well, but the twitter

crowd felt that it is necessary to customize their endorsement by replying to those tweets. It is important to note that these words appear in different frequencies when looking at retweets and likes compared to replies. These bar charts provide a useful guide for Cascades' decision makers to learn more about what triggered the different reactions to Stora Enso's tweets. This would then help Cascades' strategists to implement tactics that would appeal to the customers and shareholders of Stora Enso. Cascades could also learn from Stora Enso's themes that the twitter crowd endorsed which could then be implemented to grow their market share as well as their investor base.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.7.4 Usage of Most Common Words Over Time

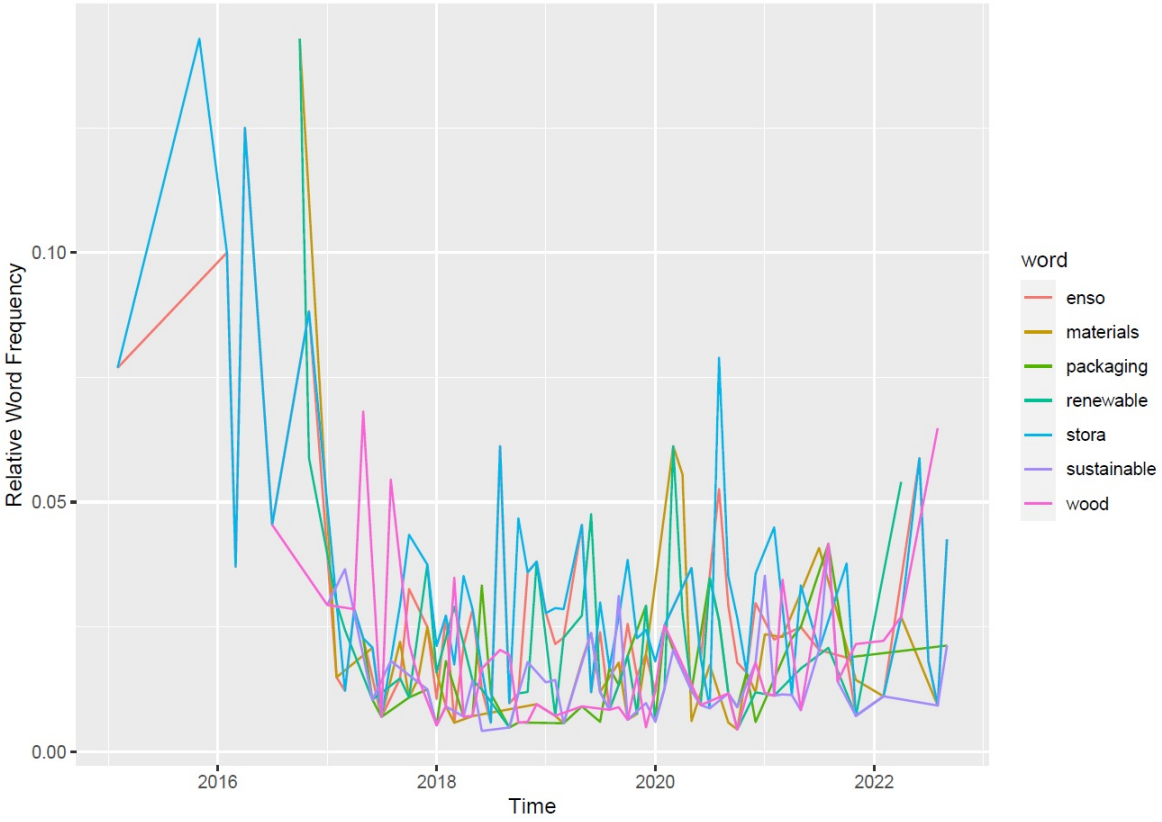


Figure 87: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection.

Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts. Figure 87 is useful to highlight the change in Stora Enso's focus in their public twitter messages over time. Besides the continuous fluctuation over time of the different topics represented by the words shown, we can see that both "wood" and "sustainable" fluctuate together. This is an indication that Stora Enso always emphasizes the sustainability in their wood product offerings.

This could be used by Cascades to track how the topics tweeted by Stora Enso evolved over time to influence their customers and shareholders in addition to identifying the upcoming trends in their competitor's behavior.

5.7.5 Structural Topic Modelling

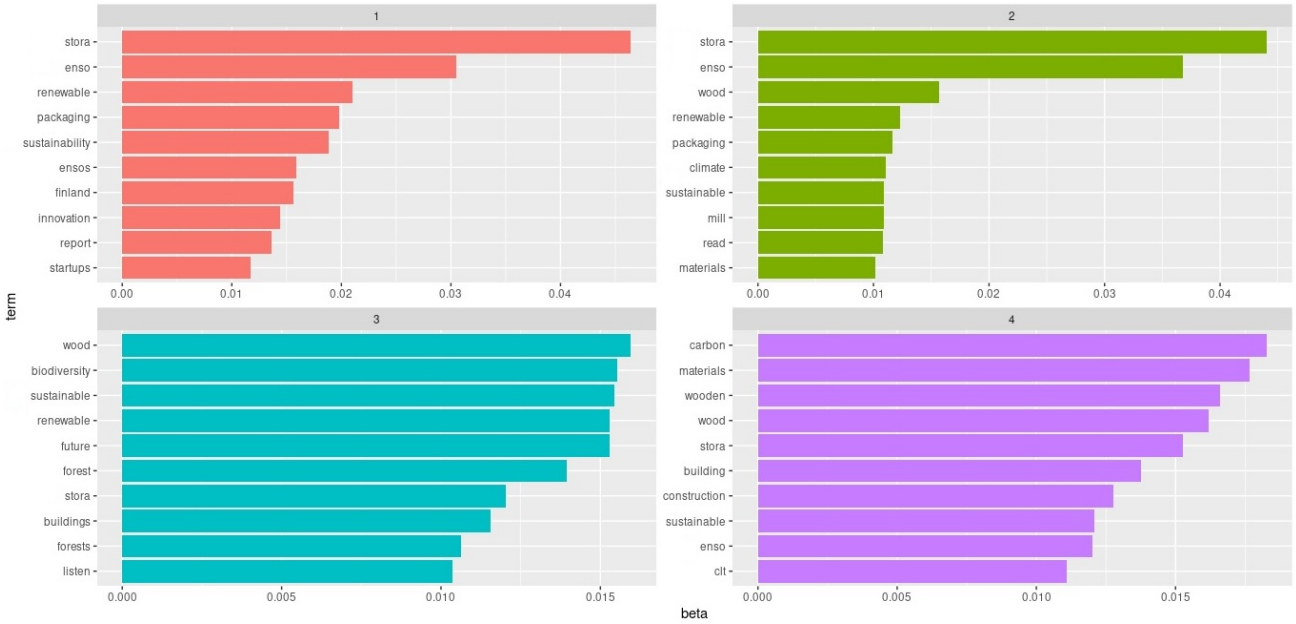


Figure 88: The optimum topic-word combination of the structural topic modelling for Stora Enso's tweets is 10 words and 4 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents Stora Enso's tweets about how they innovate and their collaborations with startups to produce their products sustainably. As for topic 2, it presents how they contribute towards fighting climate change. Regarding topic 3, Stora Enso's tweets talk about how

environment friendly their impact is. Finally, topic 4 encompasses Stora Enso's achievements in the construction industry with their Cross Laminated Timber (CLT).

It is very important to highlight that Stora Enso is not just in the pulp, paper and packaging industry, they have also pivoted into the construction industry. Thanks to the layout of the structural topic modelling, this insight was discovered.

Summing up, Stora Enso tweets about 4 topics: Innovation, fighting climate change, their general impact on the environment and their achievements in the construction industry.

5.7.6 Sentiment Analysis

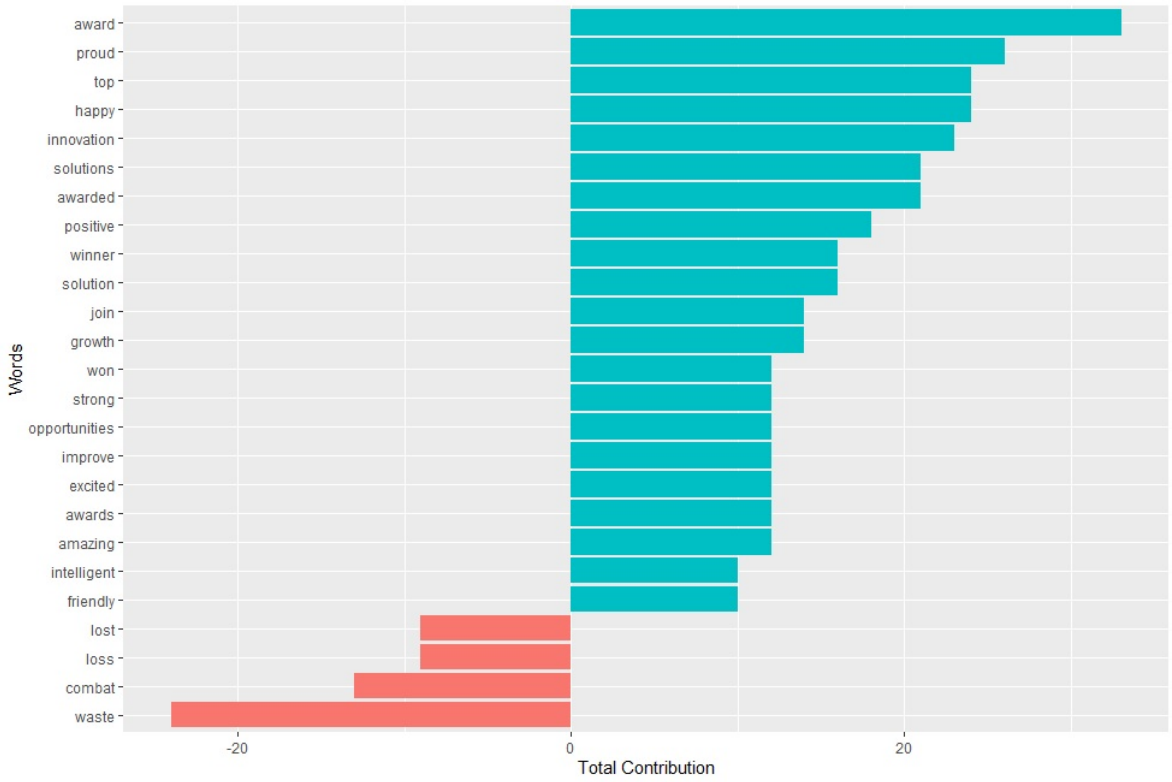


Figure 89: Sentiment analysis for Stora Enso's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive or negative sentiments. The total contribution is evaluated as the afinn score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Stora Enso. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect Stora Enso to use positive words when they publish news about the 4 topics we presented in the previous section in order to attract more

customers along with more shareholders. As expected, the top 5 unique words that have the highest contribution to Stora Enso's overall message sentiment boast about their pride and achievements, namely: award, top, proud, happy and innovation.

We can also see that the words loss, lost, combat and waste show up as the top negative contributors. A deeper drill down into the tweets containing these negative words shows that Stora Enso used the word loss to refer to the loss in nutrition due to food waste and loss in biodiversity, a true negative connotation. The word lost was also used in a negative context to talk about food waste which was also indicated as negative in our sentiment analysis. But the word combat was used in a positive way within Stora Enso's tweets to indicate how they are fighting the latter three words in addition to other adverse global effects.

We find that running a sentiment analysis on the crowds' replies to the tweets of Stora Enso would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

Our sentiment analysis also showed the presence of 2 negated words in the tweets. Thus, the presence of negation before these words flips their connotation. Overall, the total afinn sentiment score of these words does not have a sizable effect on the total sentiment score shown in the previous bar chart.

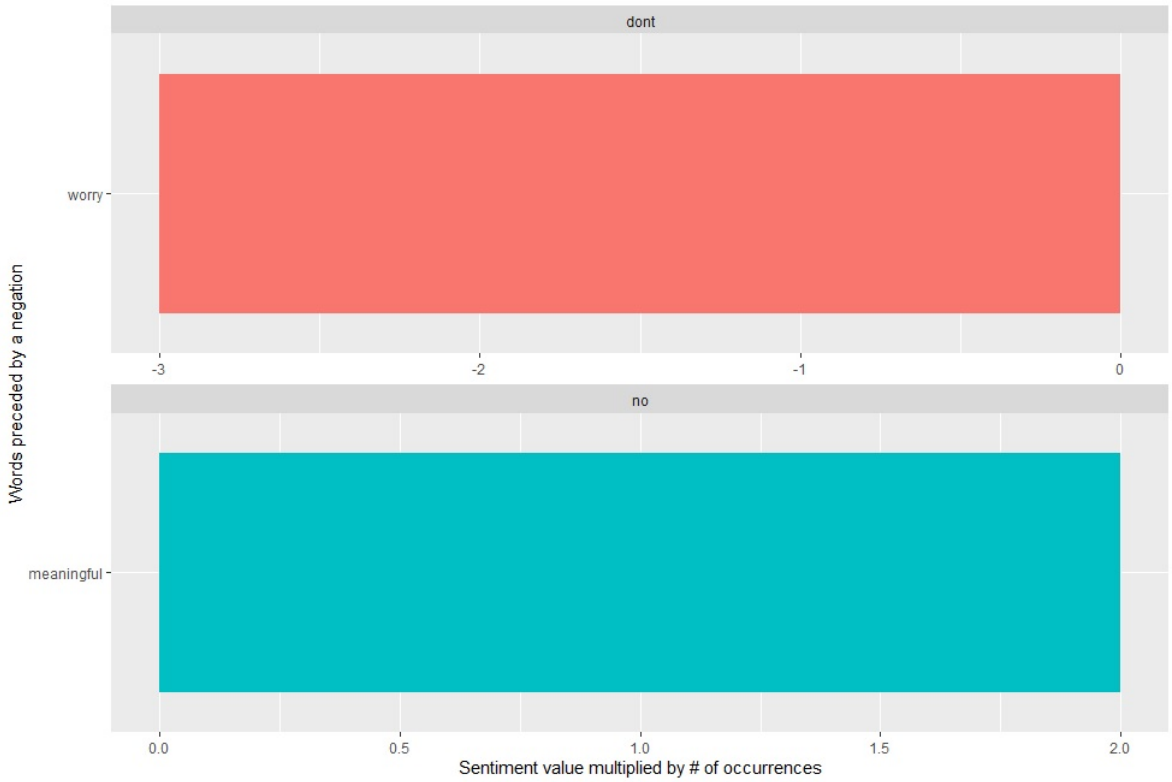


Figure 90: Sentiment analysis for Stora Enso's tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaludated as the afinn score for each word multiplied by the total number of occurrences.

5.7.7 Overall Network of Words in Tweets

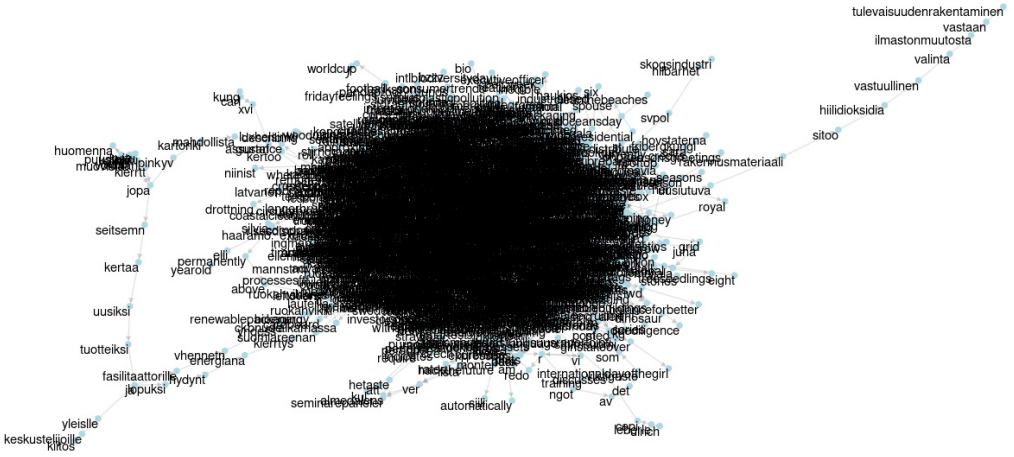


Figure 91: This directed network graph is showing the overall communication strategy of Stora Enso on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Stora Enso structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of Stora Enso. It is interesting to see how different topics start far away from the dense part of the network until they converge into that dense center. A closer look at these distant word trains indicate that they are Finnish words tweeted on the English account of Stora Enso. Nevertheless, their convergence to the dense center of the network shows how all of Stora Enso's practices tie back to their values or mission regardless of the language.

This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd while also converging to their values and mission. This also helps Cascades to fill up any structural holes in their communication and to uncover what Stora Enso talks about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.7.8 Network of Most Common Used Pairs of Words

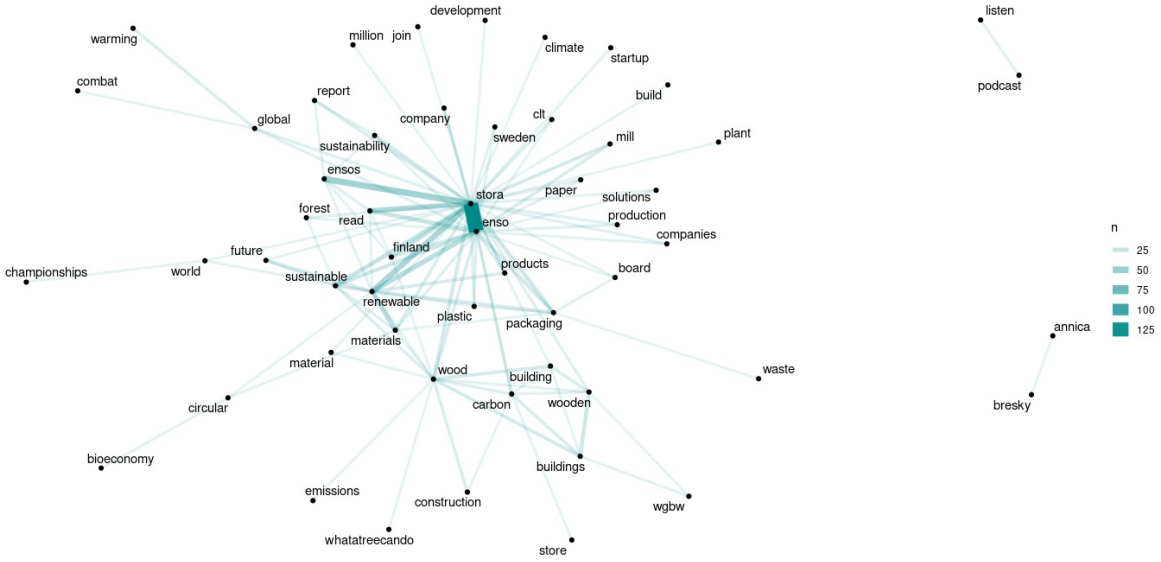


Figure 92: This network graph is showing the most common usage of the pair of words tweeted by Stora Enso. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that Stora Enso tweets about. We can clearly see the presence of several quasi star shaped graphs, albeit in different sizes. Drilling deeper into this graph shows how the different small worlds (Goyal, Van Der Leij, and Moraga-González 2006; Uzzi and Spiro 2005; Watts and Strogatz 1998) are connected to each other to form the innovative business models that Stora Enso implements, including their pivot from pulp and paper industry, to the circular economy, to the construction industry.

It would be useful for Cascades to drill down further into Stora Enso's operations and map out the different virtuous cycles in their business model in order to identify opportunities that will help them gain a competitive advantage in their geographic area or beyond.

5.8 TAPPI

Tappi (`@tappitweets`) present themselves on twitter as the international center of excellence for forest products, pulp, paper, packaging and related industries. They are based out of Peachtree Corners (Georgia) and they joined Twitter in March 2009. The tweets we were able to collect for `@tappitweets` span dates from December 2009 till November 2022.

5.8.1 Exploratory Data Analysis

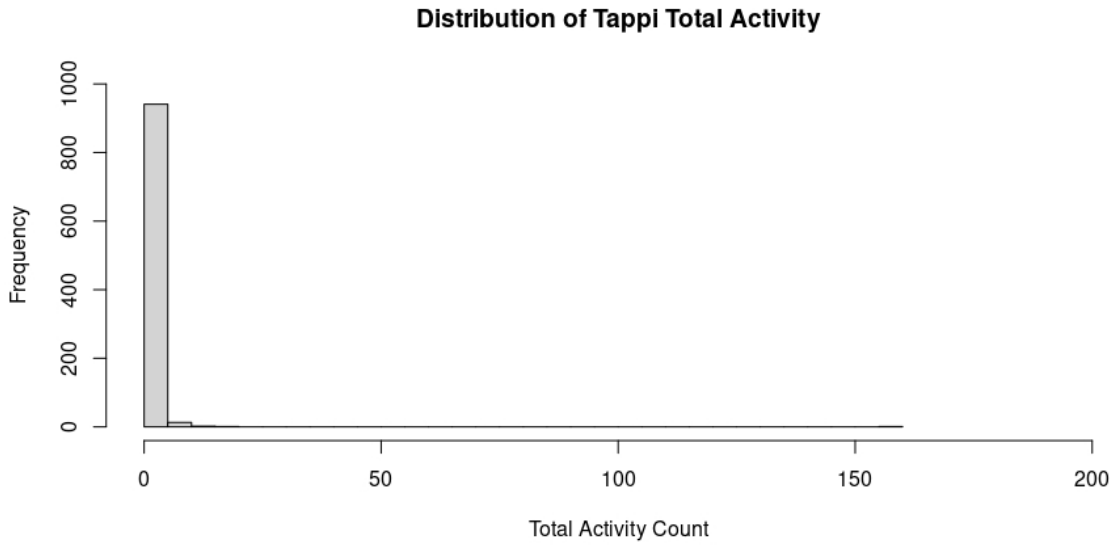


Figure 93: Distribution of total activity counts for the different tweets by Tappi

Looking at the Twitter activity distribution of Tappi, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by Tappi that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 2 for Tappi, we will take all the tweets whose total activity is larger than 2. This brings down the total number of tweets by Tappi from 958 to 142 tweets for analysis.

5.8.2 Identification and Classification of Spikes

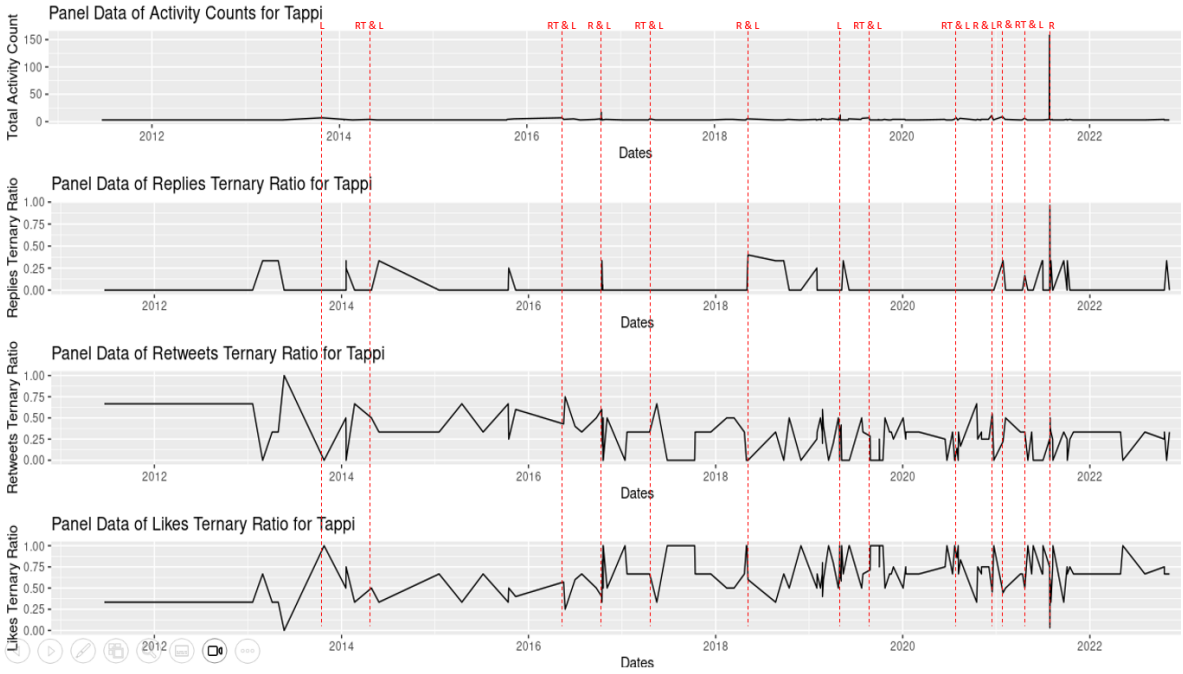


Figure 94: Graphs showing spikes in total activity for Tappi's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, Tappi presented only a few hot topics to the twitter crowd per the spikes in total activity count over the whole period of analysis. Another point to mention is how Tappi was not able at all to engage the twitter crowd except for those few spikes shown in total activity. Correlating all 4 graphs together, the majority of tweets attracted different combinations of replies, retweets and likes with just 2 tweets strictly attracting likes and the hottest tweet strictly attracting replies. This tells us that the words attracting retweets will be similar, if not identical, to the words in the tweets that attracted likes contrary to what we expect for replies. This overview of engagements tells us that the twitter crowd does not find the tweets made by Tappi as generally interesting and as a result, they do not engage with them. But it would be useful to drill down into that tweet that attracted a very large number of replies relative to retweets and likes. This could be translated into Tappi tweeting about a topic that is considered as very hot by the twitter crowd, who in turn strongly disagrees with the context of their tweet. Indeed, a deeper drill down into this particular tweet shows that Tappi announced they will be hosting 2 keynote speakers, David Sewell and Pete Watson, to speak at one of their biggest events - SuperCorrExpo 2021. Further drill down into the contents and replies to this tweet indicate

that one of the keynote speaker, Pete Watson, was the CEO of a packing company called Greif that has large operations in Russia. Consequently, the twitter crowd was urging Pete Watson to stop supporting Russia by stopping all of Greif's operations there.

This graph and analysis is a another good example of how Cascades' strategy room could benefit from the exploratory process using data science by identifying time intervals that deserve a drill down. This graph also provides a strong proof to the points mentioned in Joshua Minot's paper that replies indicate a controversial or a disagreed with topic while retweets and likes indicate an endorsed topic by the twitter crowd (Minot 2022). What follows will help us understand which words attracted each type of reaction over the whole period of analysis.

5.8.3 Frequency of Most Common Words per Reaction

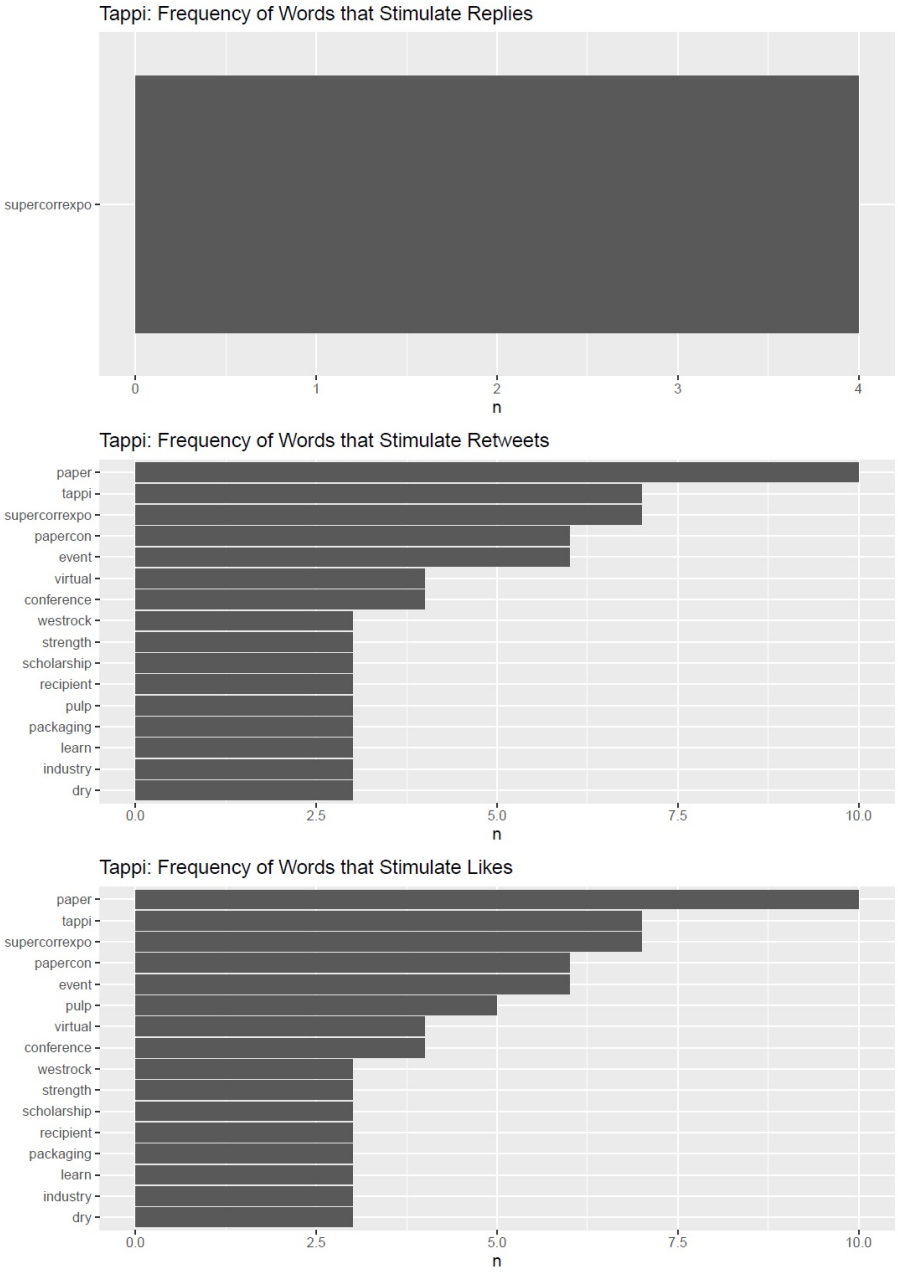


Figure 95: Frequency of words in the tweets of Tappi on hot topics

As shown in the figure above, the word SuperCorrExpo dominated the most frequent words that attracted replies due to our explanation in the previous subsection. We can also see that our expectations were true regarding the list of words attracting retweets and replies being identical. Besides that, the top words that attracted retweets and likes were revolving around the conferences and events held by Tappi and the pulp and paper topics they discuss during these events.

These bar charts provide a useful guide for Cascades' decision makers to learn more about what topics triggered the different reactions to Tappi's tweets and why these reactions were triggered. It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.8.4 Usage of Most Common Words Over Time

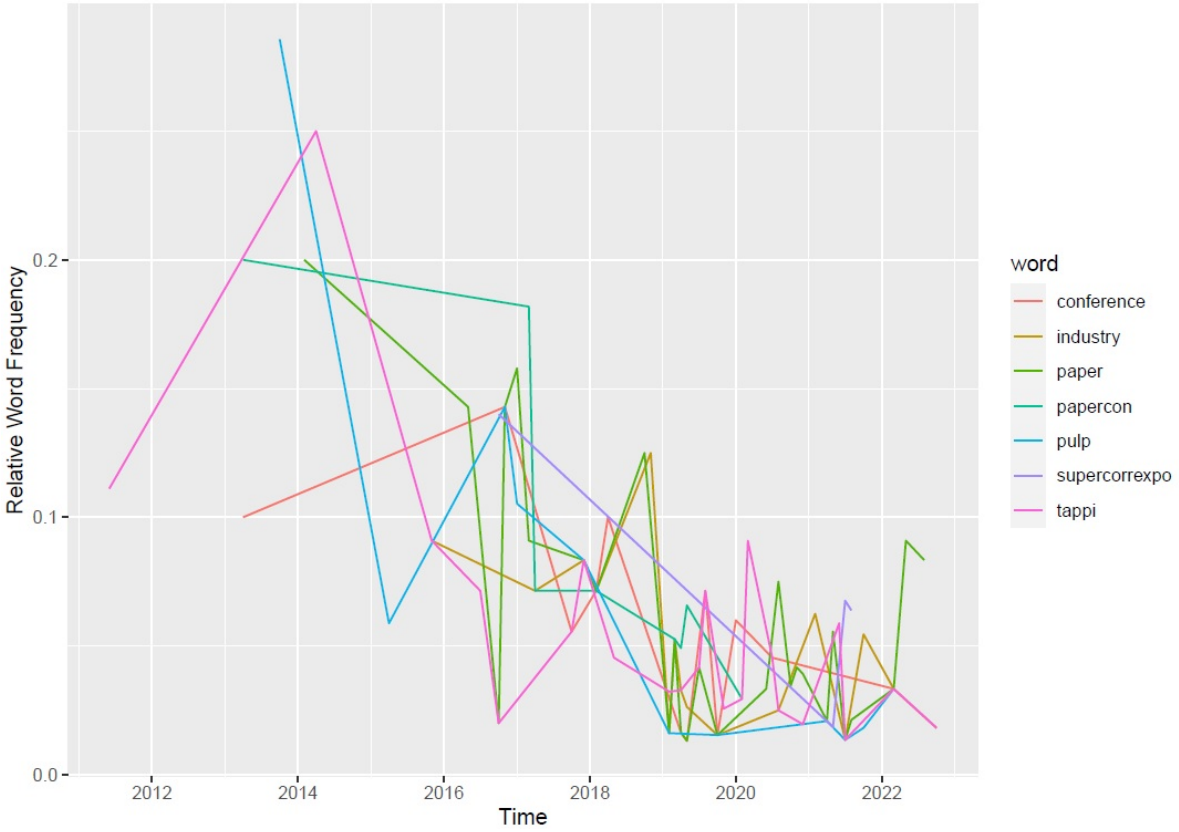


Figure 96: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts. Figure 96 is useful to highlight the change in Tappi's focus in their public twitter messages over time.

Following our synthesis in the previous subsections, it can be seen how the relative frequency of SuperCorrExpo drastically dropped after mid 2020. This particular drop could be correlated with the time period when Tappi started receiving the reactions of the twitter crowd to their tweet about the presence of Pete Watson as a keynote speaker in order to market SuperCorrExpo 2021.

What we mentioned here is a small example of how Cascades could track and analyse the topics tweeted by Tappi over time to influence their customers and shareholders in addition to identifying the upcoming trends in the pulp and paper industry.

5.8.5 Structural Topic Modelling

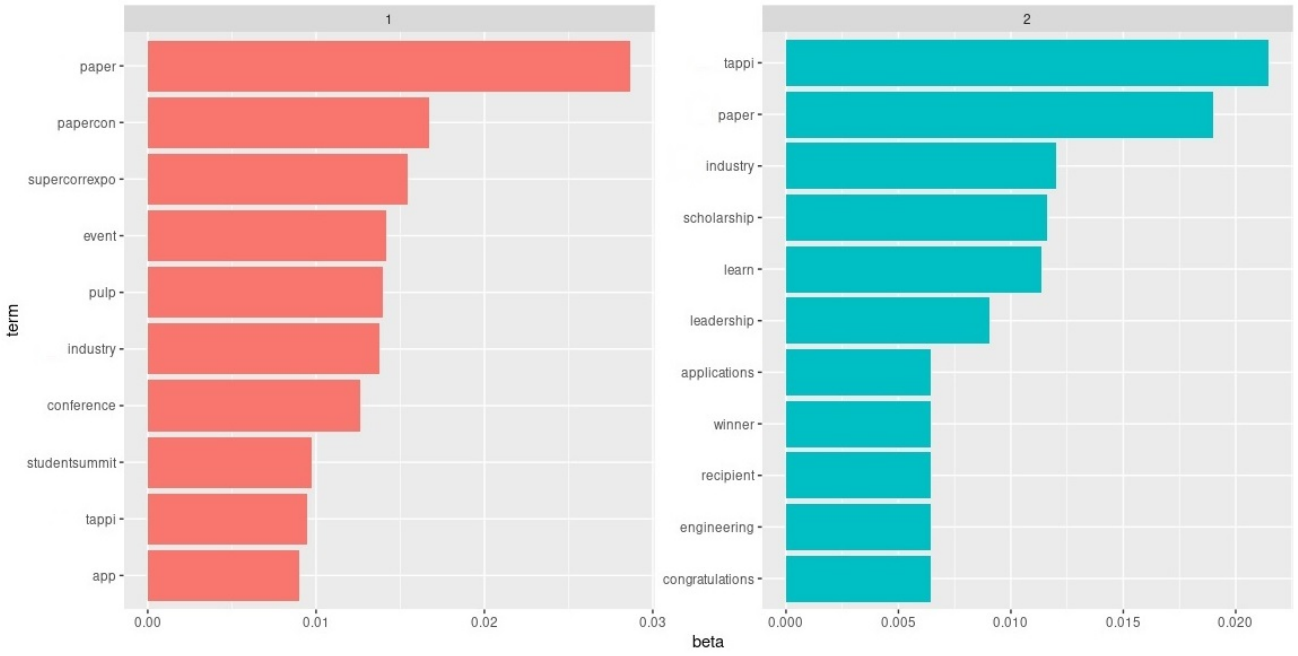


Figure 97: The optimum topic-word combination of the structural topic modelling for Tappi's tweets is 10 words and 2 topics

Knowing that Tappi is equivalent to CIRANO (CIRANO 2023) for the pulp, paper and packaging industry and based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents Tappi's tweets about their different events and conferences that they hold for their members to network. As for topic 2, it presents the awards given to their members or event participants.

Summing up, Tappi tweets about 2 topics: Their events and their awards.

5.8.6 Sentiment Analysis

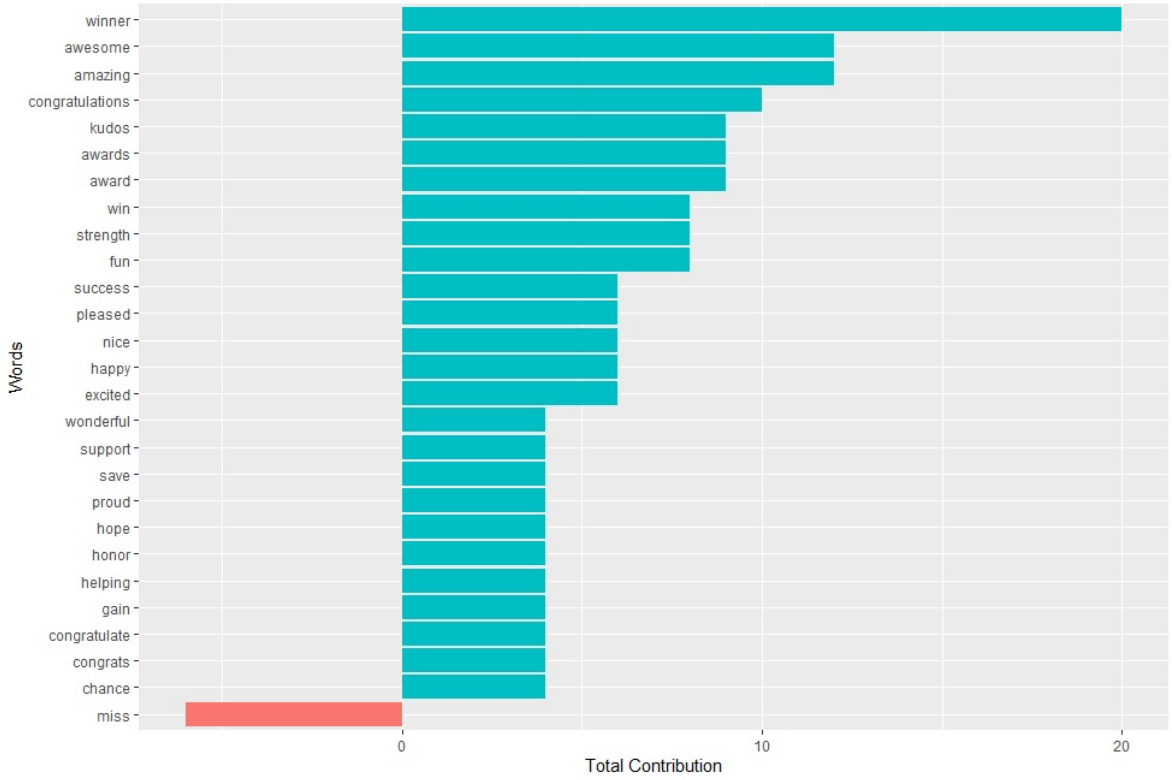


Figure 98: Sentiment analysis for Tappi's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by Tappi. Being a center of excellence that connects different players in the pulp and paper industry, we would expect them to use positive words when they tweet about their events. As expected, the top 5 unique words that have the highest contribution to Tappi's overall message sentiment include winner, awesome, congratulations, awards and strength.

We can also see that the word "miss" shows up as the top negative contributor. A deeper drill down into the tweets containing this negative word shows that it was always preceded by the negation "dont" as shown in figure 99 to urge the twitter crowd not to miss their events. Hence, a positive context after all.

We find that running a sentiment analysis on the crowds' replies to the tweets of Tappi would be very useful for Cascades' decision makers, especially if the said tweets garner a very large number of replies such as the tweet about SuperCorrExpo 2021. Employing data science to

analyse these replies would save lots of time and resources to go through and analyse them manually. Hence, the efficient exploration we referred to previously.

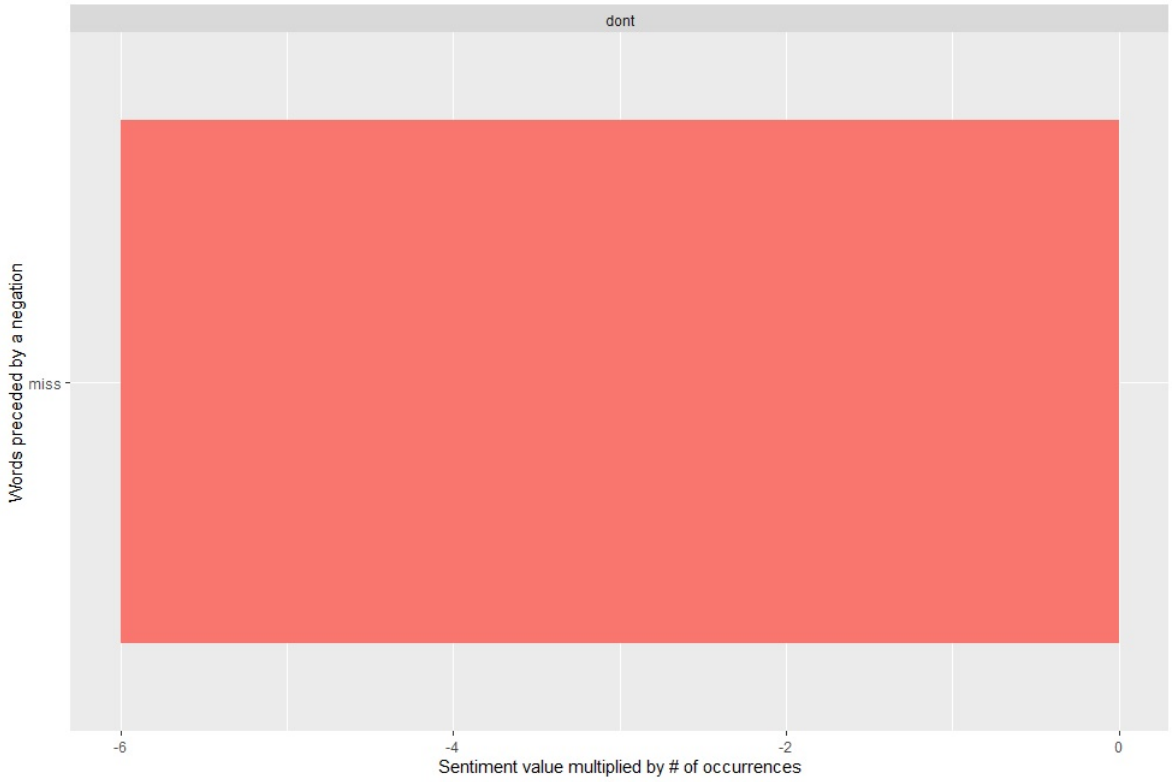


Figure 99: Sentiment analysis for Tappi's tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaludated as the afinn score for each word multiplied by the total number of occurrences

5.8.7 Overall Network of Words in Tweets

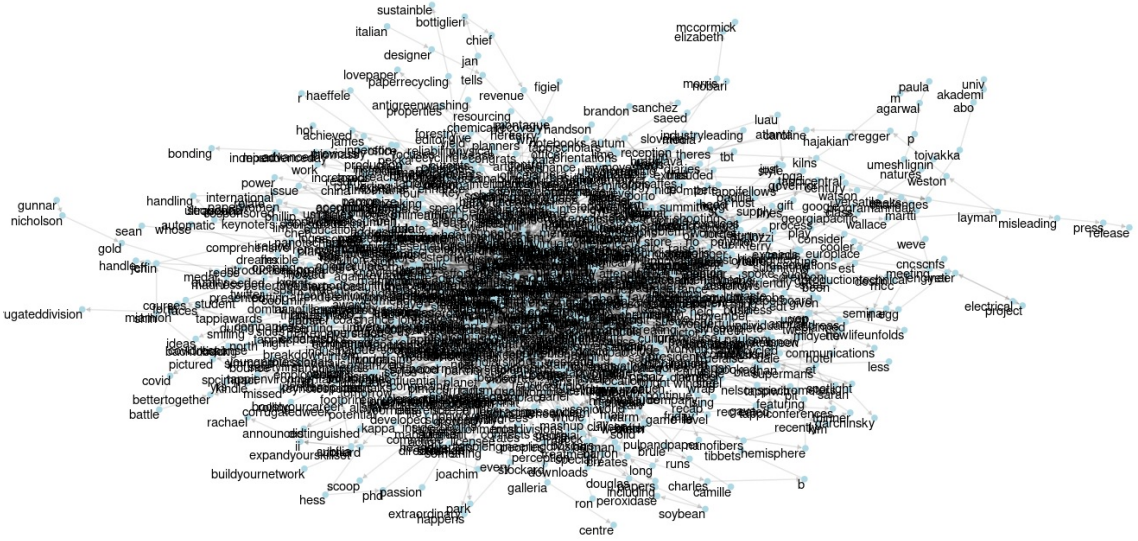


Figure 100: This directed network graph is showing the overall communication strategy of Tappi on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Tappi structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of Tappi. This provides a useful guide for Cascades' decision makers to learn from the topics discussed in Tappi's tweets and structure competitive tweets that appeal to the twitter crowd.

5.8.8 Network of Most Common Used Pairs of Words

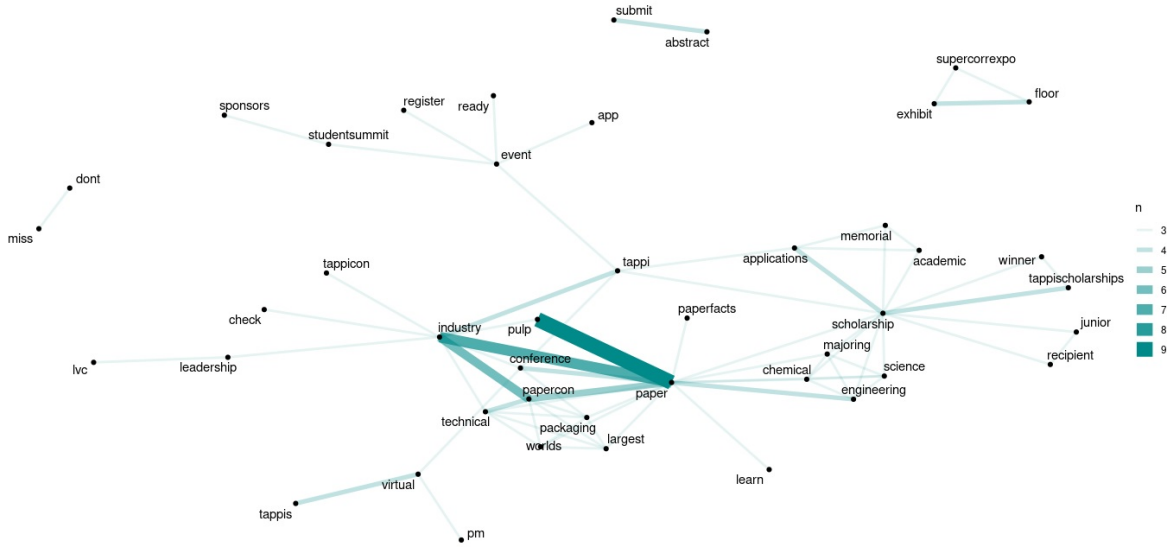


Figure 101: This network graph is showing the most common usage of the pair of words tweeted by Tappi. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases

The network graph above is another way to map out the topics that Tappi tweets about. We can clearly see the presence of several quasi star shaped graphs, albeit in different sizes. Drilling deeper into this graph shows how the different small worlds (Goyal, Van Der Leij, and Moraga-González 2006; Uzzi and Spiro 2005; Watts and Strogatz 1998) are connected to each other to form the interconnected, yet different topics.

Even though there doesn't seem to be any insights shown in figure 101 above, it would still be useful for Cascades to drill down into different topics to see how they are interconnected before relating this graph to the other network graphs we presented.

5.9 UPM Global

UPM Global (`@upmglobal`) present themselves on twitter as the Biofore company, their account keeps the twitter crowd up to date with the latest UPM news and updates (#UPM #biofore #beyondfossils). While they also have another twitter account called (`@upmsuomi`) that tweets their updates in Finnish, it was eliminated throughout our filtering process since it does not tweets in English language which would prevent us from performing Natural Language Processing (NLP). They are based out of Helsinki (Finland) and they joined Twitter in August 2009. The tweets we were able to collect for `@upmglobal` span dates from December 2009 till November 2022.

5.9.1 Exploratory Data Analysis

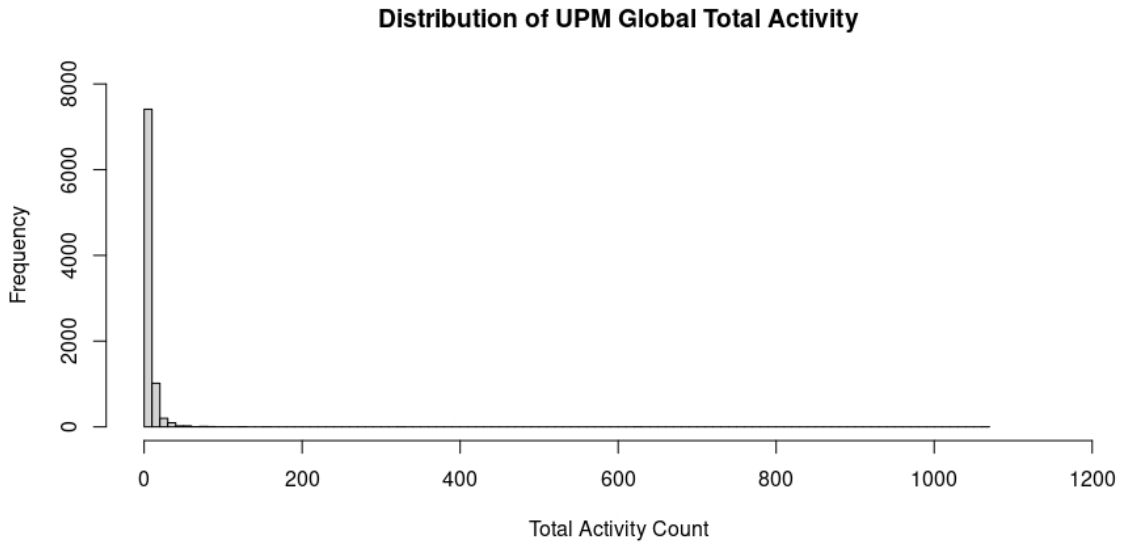


Figure 102: Distribution of total activity counts for the different tweets by UPM Global

Looking at the Twitter activity distribution of UPM Global, it is non-normal and heavily skewed to the right per the diagram above. Hence, finding the mean of the total activity counts and taking values above the mean in the right tail would be misleading. Also, excluding the outliers or tampering in anyway with the distribution of the activity data will eliminate important information we are looking for and will prevent us from reaching our objective of discovering the topics of the tweets that were voted as most controversial or most important by the audience. For that reason we will be using the 3^{rd} quartile strategy to skim important tweets and eliminate tweets made by UPM Global that were perceived as unimportant by the twitter crowd. Knowing that the 3^{rd} quartile value of total activity count is 8 for UPM

Global, we will take all the tweets whose total activity is larger than 8. This brings down the total number of tweets by UPM Global from 8,806 to 1,892 tweets for analysis.

5.9.2 Identification and Classification of Spikes

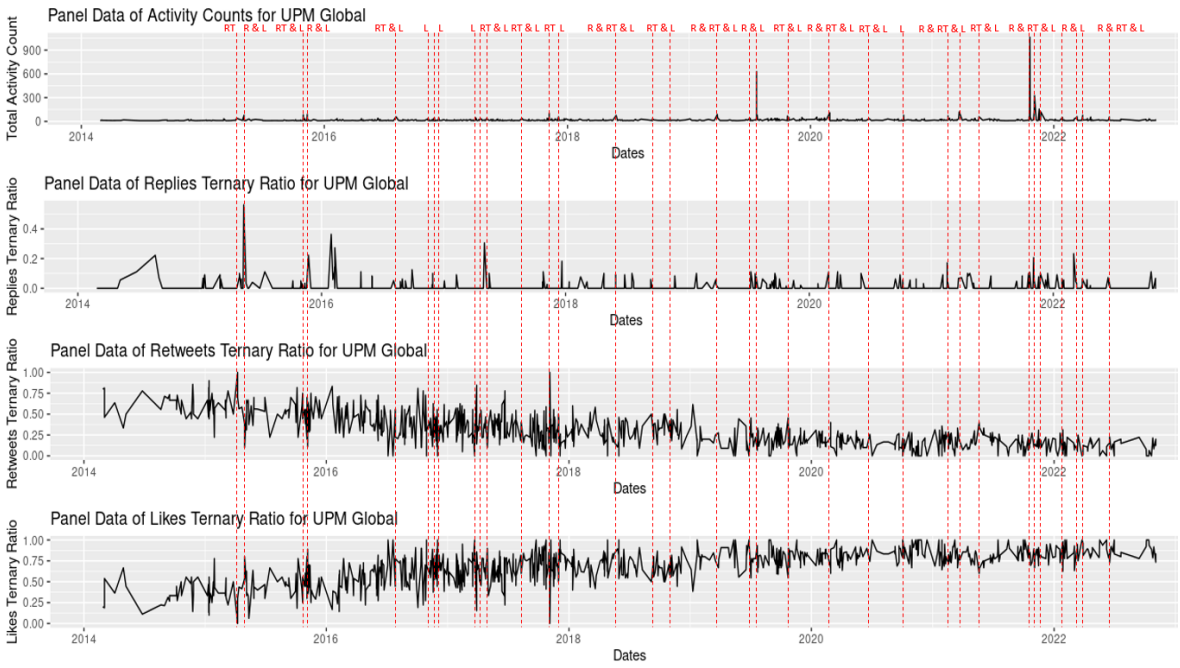


Figure 103: Graphs showing spikes in total activity for UPM Global's tweets to correlate with the spikes in ternary ratios of the different activity types. "R" represents replies, "RT" represents retweets and "L" represents likes.

As can be seen from the figure above, UPM Global presented only a few hot topics to the twitter crowd per the spikes in total activity count over the whole period of analysis. Another point to mention is how UPM Global started doing a relatively good job in keeping the twitter crowd engaged from the beginning of the year 2015. And according to the total activity count graph, the twitter crowd found that UPM Global tweeted about hotter topics during the second half of the period of analysis compared to the first half leading to the significant spikes seen. Correlating all 4 graphs together, all the tweets attracted different combinations of replies, retweets and likes except for 3 tweets attracting just likes. In addition, the lower scale of the first ternary ratio graph shows a low concentration of replies throughout the period of analysis. This overview of engagements tells us that the tweets made by UPM Global are mostly endorsed by the twitter crowd and are viewed as credible.

It would be interesting to drill down further into the contents of the 4 tweets showing significant spikes around mid year 2019 and at the end of the year 2021. These tweets

garnered a relatively much larger number of likes and retweets compared to replies and they discussed how UPM Global integrates sustainable forest management into the heart of their Finnish operations as well as their biofuels innovation.

This graph and analysis is another good example of how Cascades' strategy room could benefit from the exploratory process using data science by identifying time intervals that deserve a drill down. What follows will help us understand which words attracted each type of reaction over the whole period of analysis.

5.9.3 Frequency of Most Common Words Over Time

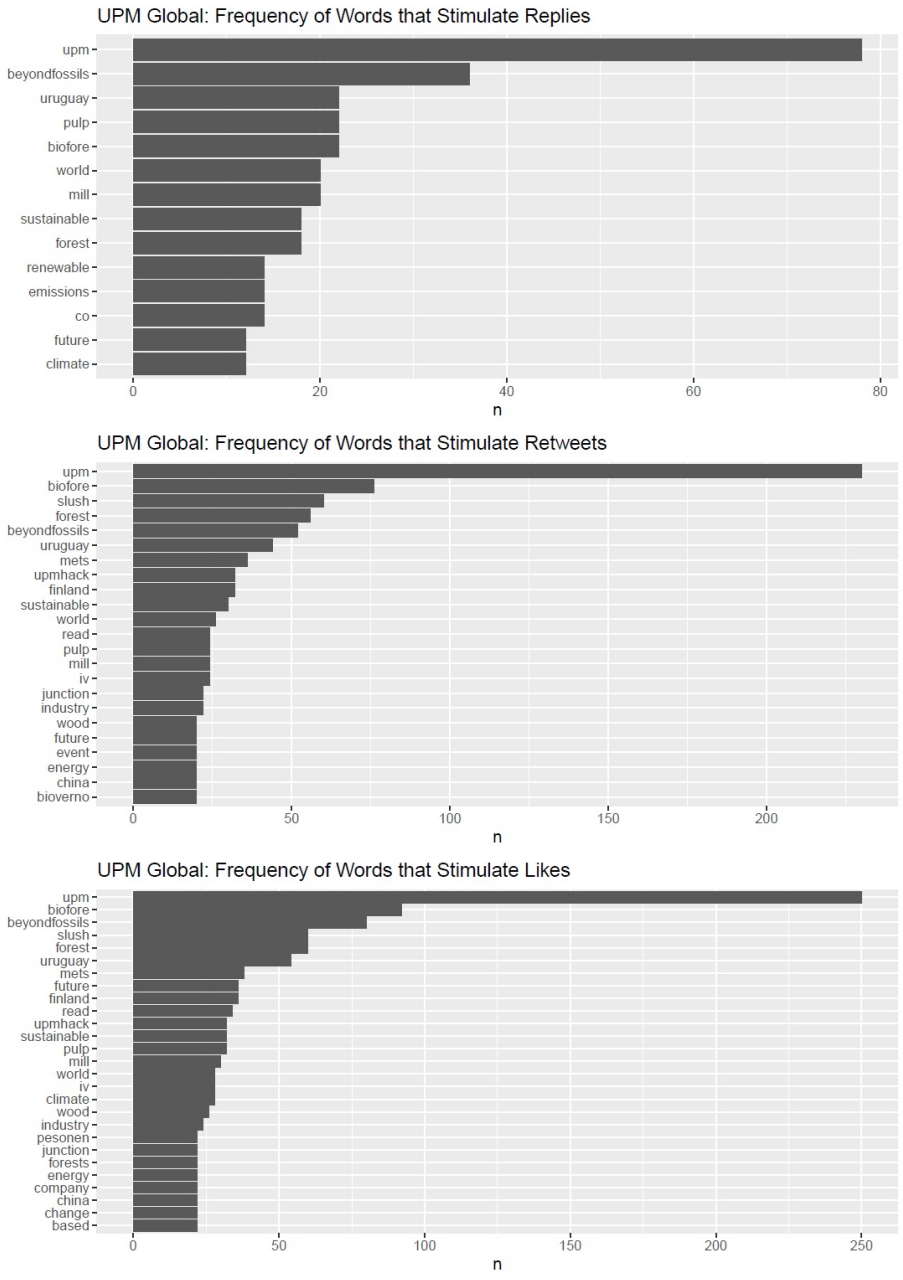


Figure 104: Frequency of words in the tweets of UPM Global on hot topics

As shown in the figure above, the words upm and beyondfossils were among the top 5 words that attracted all 3 types of reactions having the highest frequencies of all. We can also see the words attracting each type of reaction are similar but not identical. Drilling deeper into these bar charts provides a useful guide for Cascades' decision makers to learn more about what triggered the different reactions to UPM Global's tweets. This would then help Cascades' strategists to implement tactics that would appeal to the customers and shareholders of UPM Global. Cascades could also learn from UPM Global's themes that the twitter crowd endorsed

which could then be implemented to grow their market share as well as their investor base.

It is important to note that with the presence of more data, these frequency bar charts could be broken down further to separately cover consecutive time periods showing the changing tastes of the twitter crowd over time, which brings us to our following point.

5.9.4 Usage of Most Common Words Over Time

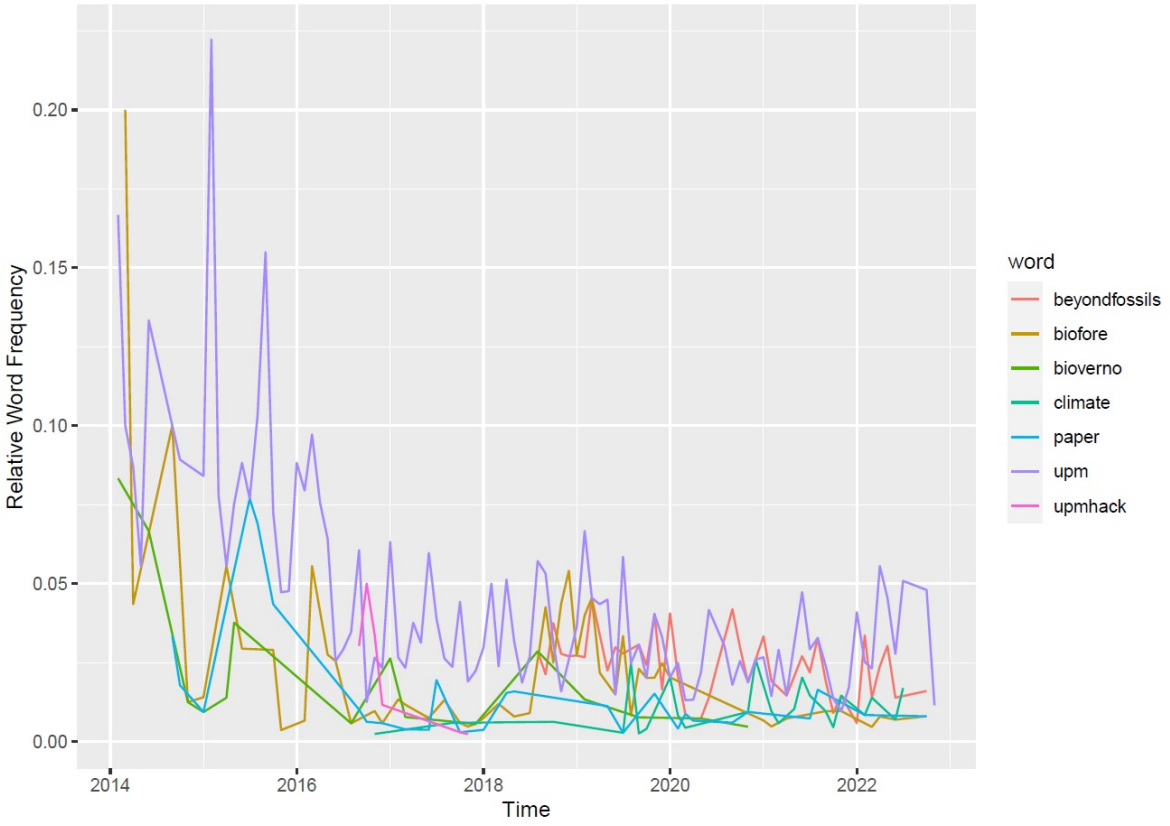


Figure 105: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals

The figure above shows the relative frequency of the most common words over time. The relative frequency is different from the absolute frequency shown in the previous subsection. Indeed, it is the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. It is important to highlight that specifying the time period depends on the average density of tweets made by the twitter account over the whole period of analysis. As a result, this measure would help us perform equitable analysis for all the different twitter accounts. This figure is useful to highlight the change in UPM Global's focus in their public twitter messages

over time. Besides the continuous fluctuation over time of the different topics represented by the words shown, we can see that upmhack emerged around mid 2016 and continued until the end of 2017 to indicate the hackathon event that UPM Global participated in to help them crowdsource innovative ideas for their different operations. We can also see that UPM Global was always tweeting about bioverno, their innovative wood-based renewable biofuel, throughout the whole period of analysis. But they launched their beyondfossils initiative on twitter around mid year 2018 which continues till the end of our analysis period in order to further strengthen their sustainability PR efforts.

This could be used by Cascades to track how the topics tweeted by UPM Global evolved over time to influence their customers and shareholders in addition to identifying the upcoming trends in their competitor's behavior.

5.9.5 Structural Topic Modelling

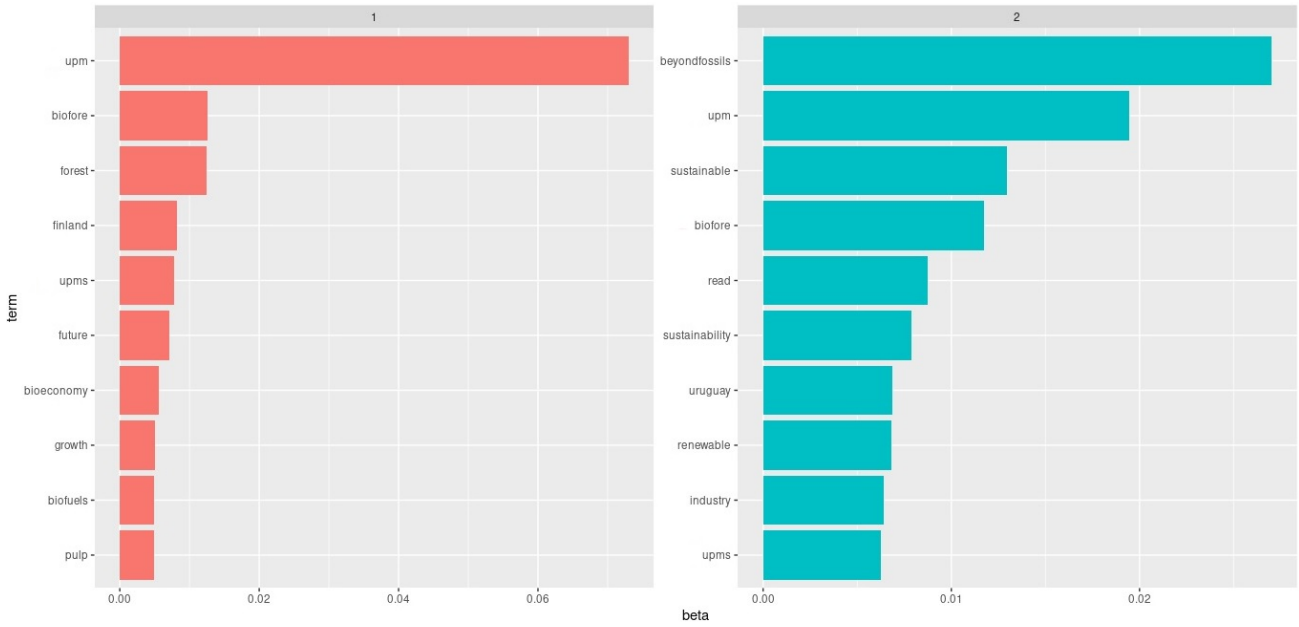


Figure 106: The optimum topic-word combination of the structural topic modelling for UPM Global's tweets is 10 words and 2 topics

Based on the per-word-per-topic probability values "beta", the combination of words in topic 1 represents UPM Global's tweets about the sustainable growth strategy they are pursuing in Finland such as the biofuels. As for topic 2, it presents their strategy in Uruguay using renewables.

Summing up, UPM Global tweets about 2 topics: Their strategy in Finland as well as their strategy in Uruguay.

It is important to highlight that UPM Global is not just in the pulp, paper and packaging industry, they have also pivoted into the energy industry by producing biofuels. Thanks to integrating structural topic modelling techniques, this insight was discovered to aid in the exploration process.

5.9.6 Sentiment Analysis

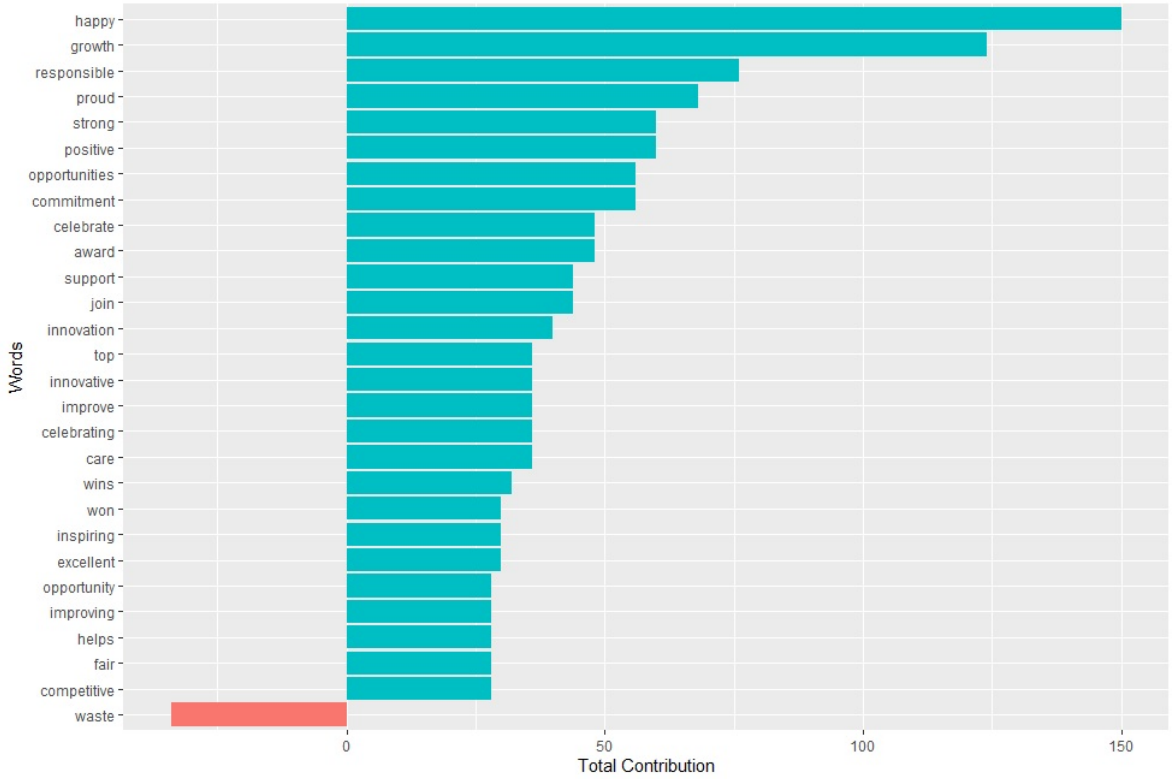


Figure 107: Sentiment analysis for UPM Global's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

To demonstrate another way how NLP can empower decision makers in the strategy room, a sentiment analysis bar chart was generated for the most common words used by UPM Global. As a for-profit pulp and paper company that cares about their public image within the competitive landscape, we would expect UPM Global to use positive words when they publish news about the 2 topics we presented in the previous section in order to attract more customers along with more shareholders. As expected, the top 5 unique words that have the highest contribution to UPM Global's overall message sentiment boast about their pride and achievements, namely: happy, growth, responsible, proud and strong.

We can also see that the word "waste" shows up as the top negative contributor. A deeper

drill down into the tweets containing the word “waste” indicates that UPM Global used it in the context to show how they reduce waste and contribute to the circular economy. We find that running a sentiment analysis on the crowds’ replies to the tweets of UPM Global would be very useful for Cascades’ decision makers, especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what their competitor is doing well and where they are lacking as decided by the twitter crowd.

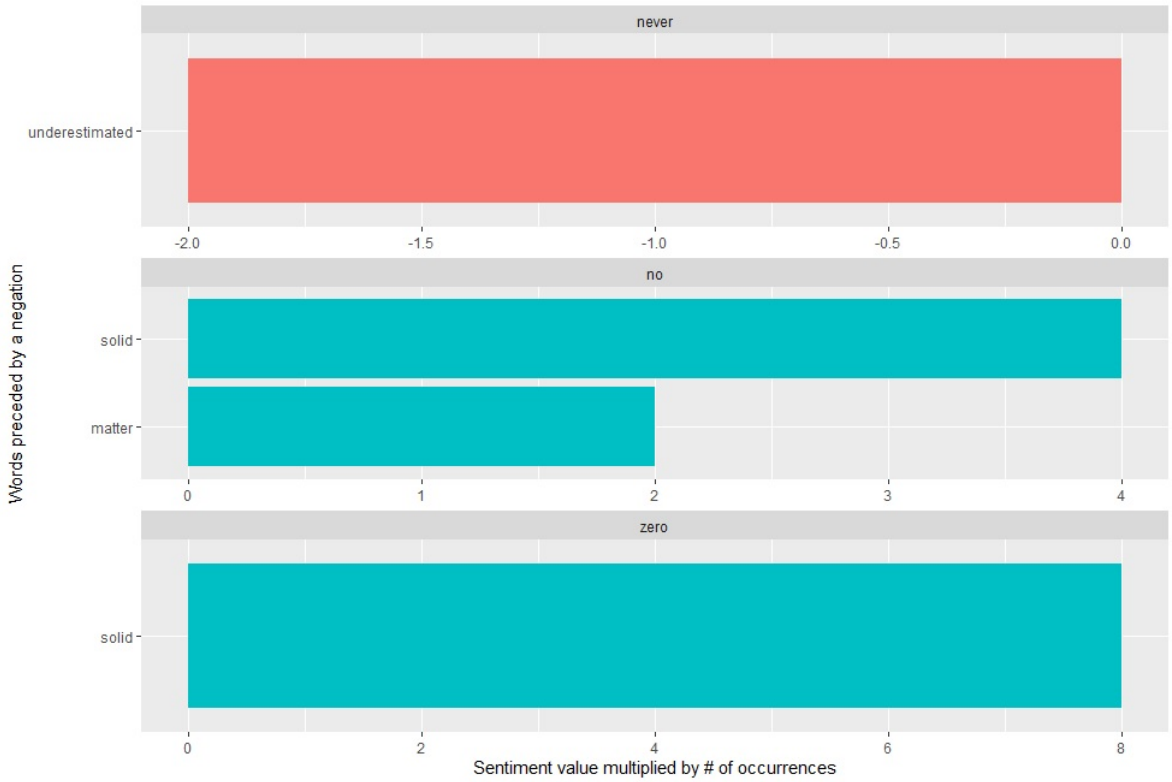


Figure 108: Sentiment analysis for UPM Global’s tweets as a whole that contain a negation. The figure shows the word(s) preceded by a negation with the highest contribution to the positive or negative sentiments. The total contribution is evaludated as the afinn score for each word multiplied by the total number of occurrences.

Our sentiment analysis also showed the presence of 3 unique negated words in the tweets. While the presence of negation before words usually flips their connotation, this case proves to be different. In the case of negating the word “solid”, it was to indicate that UPM Global aims to be solid waste free as part of their future strategy. Hence, a positive connotation by publishing their zero or no “solid” waste strategy. As for no “matter”, it was used to indicate the arrival of spring and emphasize the indifference of rain or shine. Hence, another positive connotation. As for “underestimated”, it was used to indicate that waste is never underestimated at UPM Global. Hence, a positive connotation too.

Regardless, the total afinn sentiment score of these words does not have a sizable effect on the total sentiment score shown in the previous bar chart.

5.9.7 Overall Network of Words in Tweets

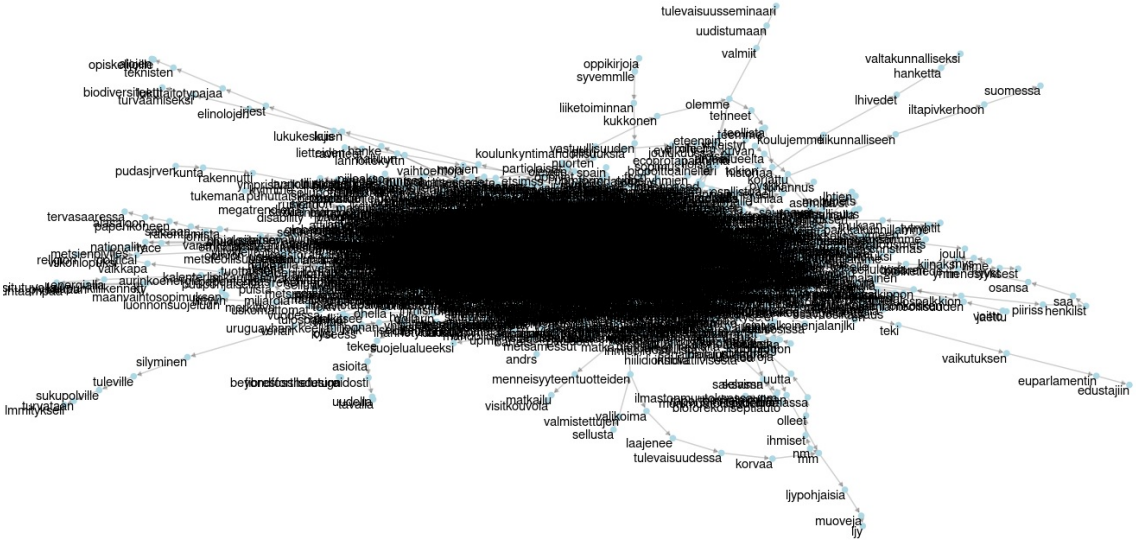


Figure 109: This directed network graph is showing the overall communication strategy of UPM Global on Twitter. The labelled nodes represent all the words used and the directed edges indicate how UPM Global structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps out the current communication strategy of UPM Global. It is interesting to see how different topics start far away from the dense part of the network until they converge into that dense center. A closer look at these distant word trains indicate that they are mostly Finnish words tweeted on the English account of UPM Global. Nevertheless, their convergence to the dense center of the network shows how all of UPM Global's practices tie back to their values or mission regardless of the language.

This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd while also converging to their values and mission. This also helps Cascades to fill up any structural holes in their business model, uncover what UPM Global talks about and how they talk about it to the public in order to get closer to their customers and shareholders.

5.9.8 Network of Most Common Used Pairs of Words

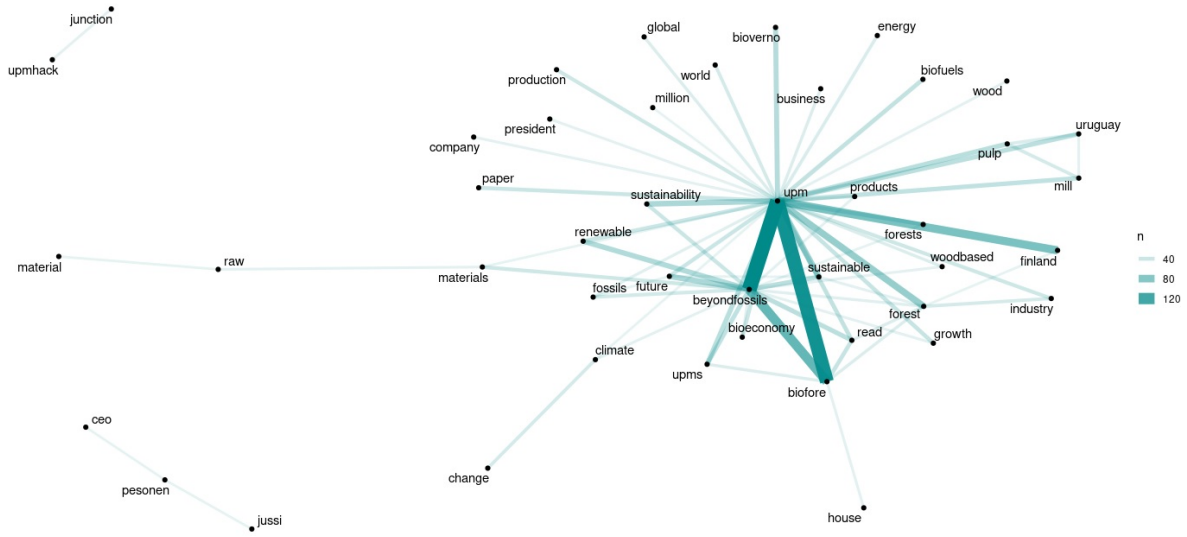


Figure 110: This network graph is showing the most common usage of the pair of words tweeted by UPM Global. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The network graph above is another way to map out the topics that UPM Global tweets about. We can clearly see the presence of several quasi star shaped graphs, albeit in different sizes. Drilling deeper into this graph shows how the different small worlds (Goyal, Van Der Leij, and Moraga-González 2006; Uzzi and Spiro 2005; Watts and Strogatz 1998) are connected to each other to form the innovative business models that UPM Global implements, including their pivot from pulp and paper industry, to the circular economy and the energy sector. It is also interesting to note how the algorithm generated graph recognized that the CEO of UPM Global is Jussi Pesonen. The graph also shows how the beyondfossils strategy is almost as central as UPM Global itself to the other topics within the graph since both words are connected to almost the same words. This indicates that the beyondfossils strategy is an integral part of the UPM brand.

It would be useful for Cascades to drill down further into UPM Global's operations and map out the different virtuous cycles in their business model in order to identify opportunities that will help them gain a competitive advantage in their geographic area or beyond.

5.10 Summary of Words Attracting Different Twitter Activities

The following gridded wordclouds provide a useful representation for comparison purposes of what practices each twitter account is involved in and where their focus lies.

5.10.1 Representation of Most Common Words Attracting Replies

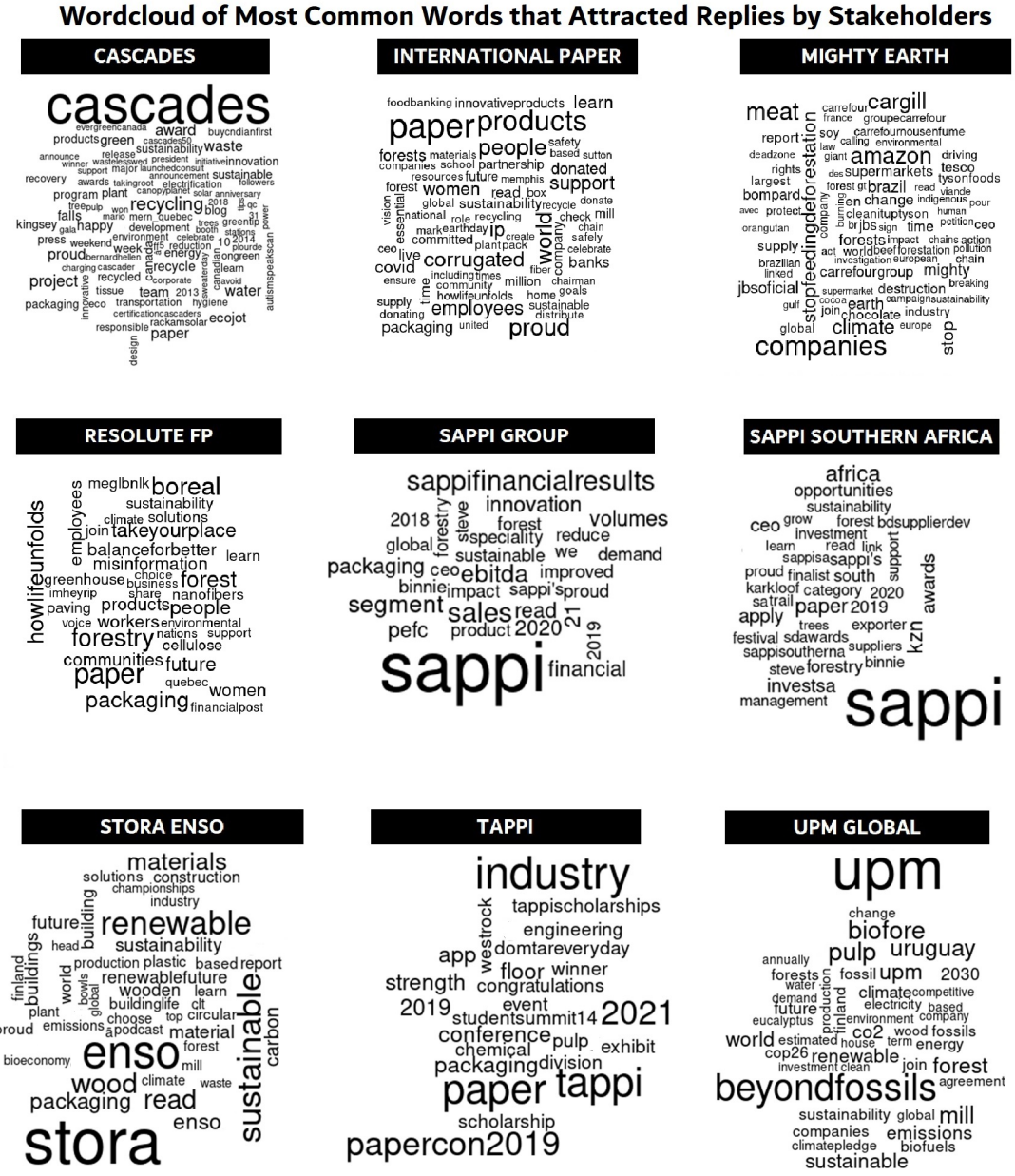


Figure 111: Wordcloud representation of the most common words tweeted by the different twitter accounts resulting in a significantly dominant replies type of activity. A larger word size represents a more commonly used word

5.10.2 Representation of Most Common Words Attracting Retweets

Figure 112: Wordcloud representation of the most common words tweeted by the different twitter accounts resulting in a significantly dominant retweets type of activity. A larger word size represents a more commonly used word

5.10.3 Representation of Most Common Words Attracting Likes

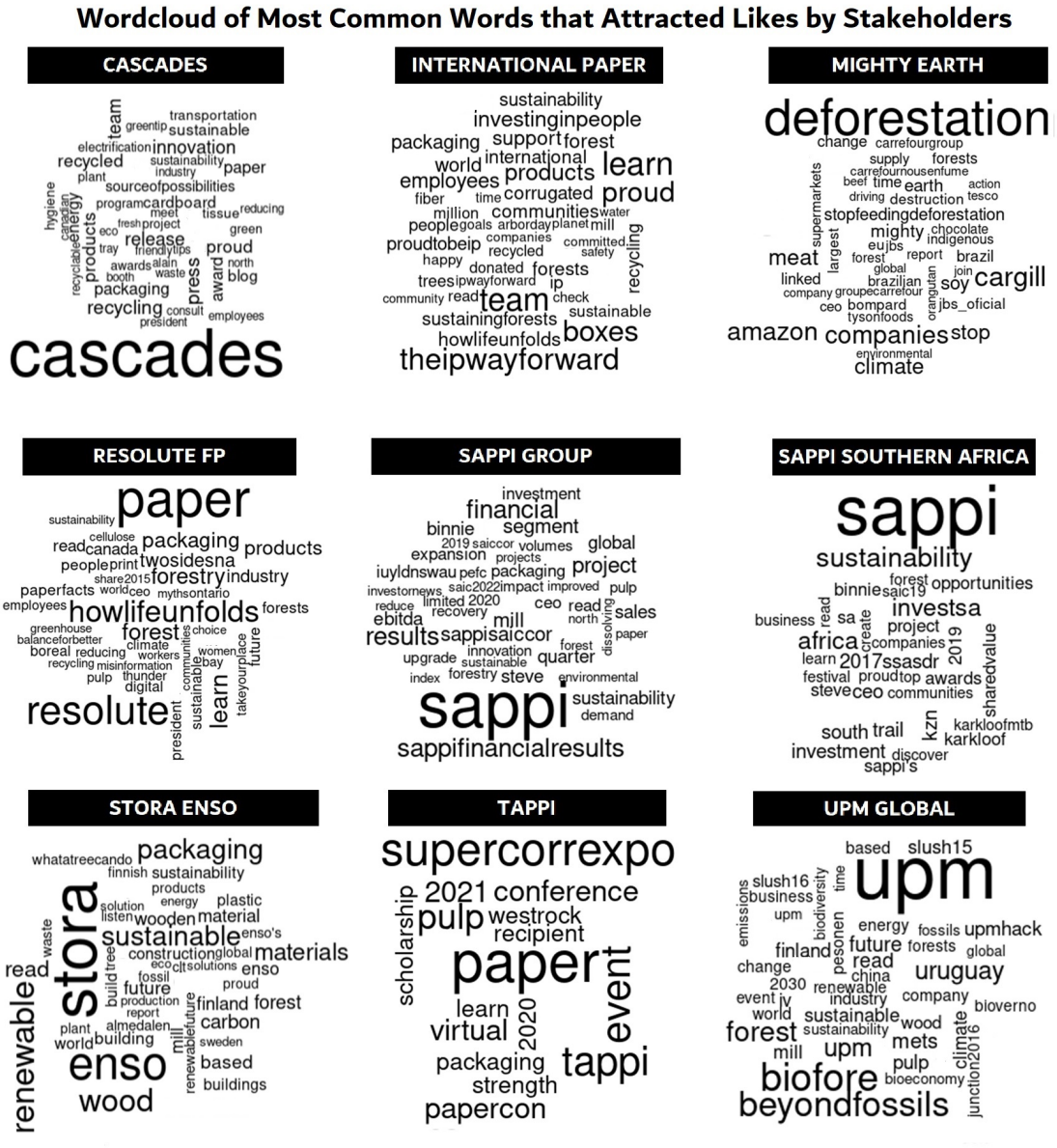


Figure 113: Wordcloud representation of the most common words tweeted by the different twitter accounts resulting in a significantly dominant likes type of activity. A larger word size represents a more commonly used word

6 Conclusion

Our goal in this dissertation was to build and test a proof concept that integrates data science into the exploration process of organisational ambidexterity. To achieve that, we examined a total of 21,189 tweets spanning dates from December 2009 to November 2022. These tweets were made by 7 of the global top 20 pulp and paper companies by turnover, 1 pulp and paper center of excellence in addition to 1 environmental lobby group. The pulp and paper industry was chosen to provide a stronger proof of concept by testing it on a niche industry where such methods were never used as far as we know. And Cascades, who is one of the 7 pulp and paper companies under analysis, was chosen randomly as the test subject in order to validate if our proof of concept achieves the goals we had set. All the steps in this dissertation were accomplished using code in R language.

Data and metadata were collected by interacting with Twitter's rest Application Programming Interface (API). This is synonymous to the search and discovery actions of the exploration process outlined in figure 23 of the Literature Review chapter. From an ethical standpoint, this data is considered as primary data and its collection falls within the ethical boundaries for two reasons: Firstly, this data was made public by their owners through tweets on the world wide web. And second, the owners of these tweets had already signed the privacy agreement provided by twitter and any use of this data naturally inherits the privacy agreement of twitter that states: "Twitter is public and Tweets are immediately viewable and searchable by anyone around the world" (T. Twitter 2020). This encouraged putting more focus on the quality of the data on which we built upon our research. Hence, our data science approach empirically addresses the heightened levels of uncertainty that we referred to in our overarching research question.

Then, algorithms were used to identify and eliminate less important tweets for each of the 9 twitter accounts as decided by the twitter crowd through their level of engagement with each tweet. The level of engagement was measured as the total activity count for each tweet, i.e. the sum of the counts of replies, retweets and likes to each tweet. Then, the level of engagement for each twitter account was graphed over time to identify the time intervals where the twitter conversations picked up, as demonstrated by spikes in total activity counts. Then, the ternary ratio methodology was employed to identify the nature of the engagement(s) towards the "spiky" tweet, i.e. replies, retweets, likes or different combinations of the three engagements. This stage marks the first, but not only step, in addressing the

heightened levels of complexities in the exploration process that we referred to in our overarching research question.

Thereafter, text mining techniques were used to analyse the resulting unstructured data, that takes the form of tweets, before producing charts showing the frequencies of the most common words that stimulated each type of engagement for each twitter account over the whole period of analysis. In addition, another graph was produced for each twitter account to examine the evolution of the most common words over time by studying their usage frequency relative to the usage frequency of the other words. As a result, this particular stage marks an important step in untangling the complexities we referred to in our research question. Thus, reducing the rigidity and increasing the flexibility embedded in the exploration process.

From there on, different Natural Language Processing (NLP) algorithms were used to further analyze the tweets made by the 9 twitter accounts. Our NLP started with structural topic modelling on each twitter account to contextualize the contents of their tweets by grouping them into an optimum number of topics made up of a mixture of words. We then continued our NLP with sentiment analysis to complement the results from structural topic modelling and evaluate whether the tweets made by each twitter account are positive or negative in addition to how positive or negative using the AFINN lexicon. This particular stage is also key in moderating the complexities associated with the exploration process. Another way to appreciate the value of our data science approach is to imagine doing these activities manually.

After that, we mapped the communication strategy of each twitter account by producing a directed network graph showing the interconnection between all the words used in their tweets, followed by an undirected network graph showing the interconnection between the most common pairs of words used in their tweets. This stage also characterizes how our data science approach tackles the heightened competitiveness that we referred to in our research question from the communication strategy point of view. Besides, this stage complements the previous stages in order to identify pivot opportunities and build upon the innovations of competitors.

Finally, we produced a wordcloud grid of the most common words for each type of twitter engagement covering the 9 twitter accounts under analysis.

Our aim through our proof of concept was to find out if data science can help reduce uncertainty, uncover pivot opportunities and augment the decision making abilities of the strategy room. Indeed, data science empowered us to collect huge amounts of data by simultaneously listening to the twitter conversations of numerous stakeholders in the pulp and

paper industry. Hence, reducing uncertainty by having all the productive knowledge available in the industry, an advantage that is not technically possible without data science. In addition, data science enabled us to filter through all of these conversations to seamlessly detect hot topics as decided by the stakeholders. Even though the exploration process for hot topic detection is possible without employing data science, the manual process is long, cumbersome, uncertain and expensive. Thus, very resource intensive. Furthermore, data science proved its ability to streamline and organize the large amounts of data into powerful visuals that could be used to produce insights and uncover new business opportunities. Consequently, empowering the strategy room to make decisions with a higher level of confidence. The following subsection provides few examples to prove that our results largely align with the intention of our conceptual framework, i.e facilitate the exploration process of organisational ambidexterity to reap all its benefits without sacrificing large amounts of resources. The subsection after that will discuss the limitations of our methodology. Finally, we will propose future research avenues and end this dissertation where we started.

6.1 Notable Findings

To effectively showcase the efficacy of our conceptual framework, it's prudent to structure this section into distinct subsections. Each subsection will correspond to a specific step outlined in our methodology. Within each of these subsections, we will elucidate the respective step by presenting a pertinent example. This segmented approach not only ensures clarity and coherence but also provides readers with a step-by-step walkthrough of our methodology, reinforcing its validity and applicability.

6.1.1 Exploratory Data Analysis

Correlating the distribution above with our knowledge of the type of data, i.e distinct counts of engagements to tweets, we were able to design an elimination strategy of the less important tweets by determining how many engagements make a tweet important. In addition, it helps us level the playing field by having a standardized elimination strategy that fits all the different twitter accounts (3^{rd} quartile value of total activity for all tweets) rather than a unified count of total activity. Moreover, it ensures we do not mask or dilute the contents of the important topics by including the large number of unimportant tweets shown on the left side. Hence, providing a filtering mechanism in our exploration process which is very resource intensive when done manually.

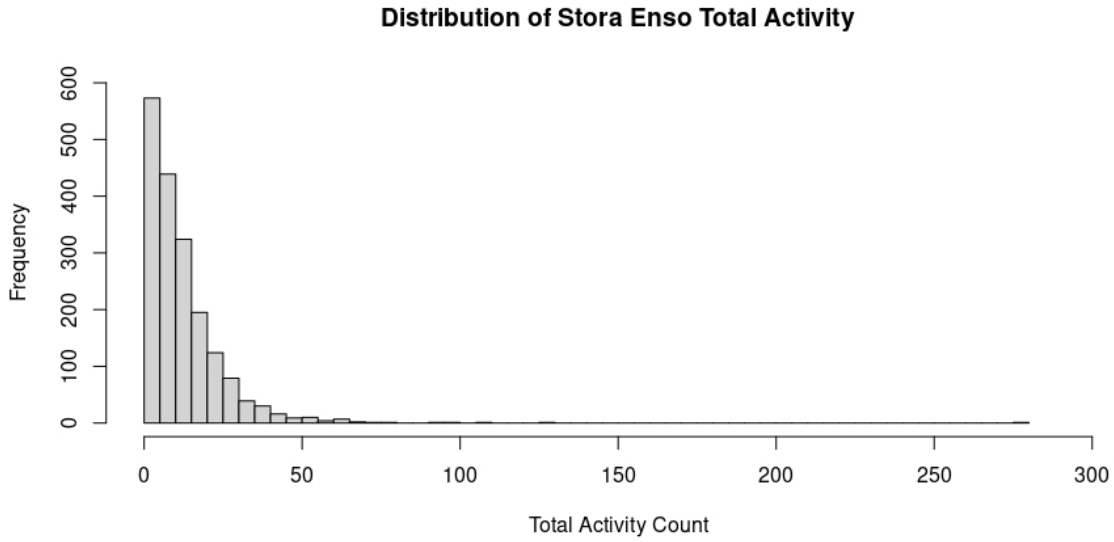


Figure 114: Distribution of total activity counts for the different tweets by Stora Enso

6.1.2 Identification and Classification of Spikes

The total activity count graph above shows a major spike around mid year 2021. Therefore, the twitter crowd finds that this tweet discusses a hot topic, which motivates them to engage heavily with it. When correlated with the 3 ternary ratio graphs, we see a spike in replies compared to zero or negligible values of ternary ratio for retweets and likes. This implies that the twitter crowd strongly disagrees with this topic, or at best finds that a simple like or retweet is not enough which motivates them to better express how they feel about this tweet in writing. In a traditional exploration process, through focus groups for example, the two findings in the graph above could potentially be missed depending on the design of the study focus group, the level of objectivity, the sample characteristics, the resources of time, money and personnel available... etc. Whereas the data science powered exploration process depicted here eliminates the difficulties we mentioned to a far extent. As a result, decision makers can immediately investigate that "spiky" tweet to discover that the stakeholders are against doing business in Russia as an act of solidarity with the people of Ukraine. Then, they would craft strategies accordingly.

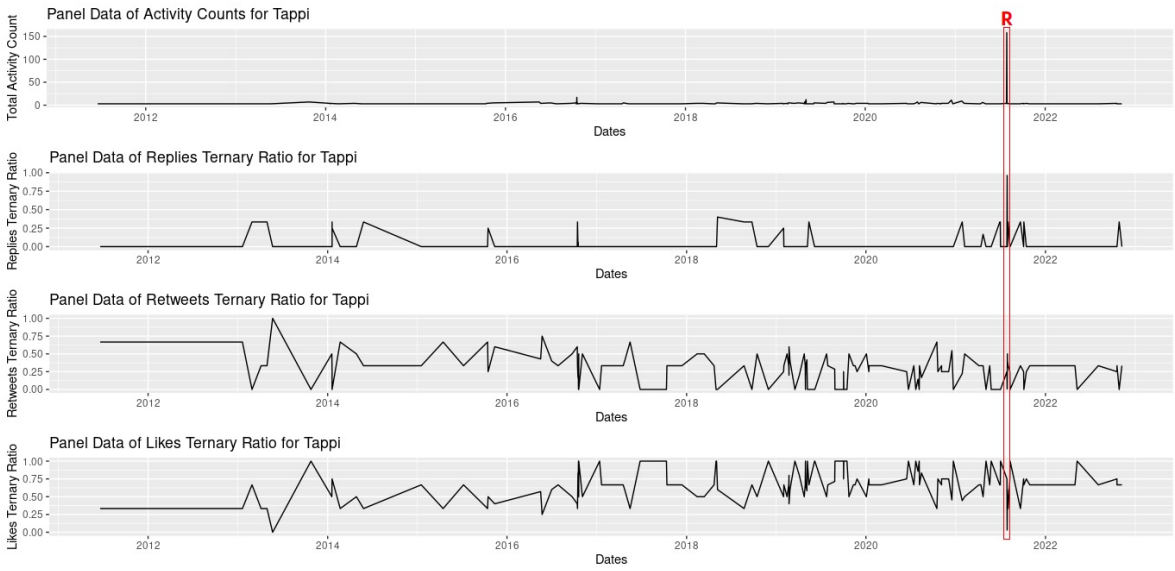


Figure 115: Graph showing spikes in total activity for Tappi's tweets to correlate with the spikes in ternary ratios of the different activity types. The "R" above the spike in total activity indicates that replies largely dominated the types of engagements towards this tweet.

6.1.3 Frequency of Most Common Words per Reaction

The per-word-per-reaction frequency bar charts shown above indicate that the word "corrugated" attracted more replies than retweets or likes. Further drill down into this intriguing result revealed that different employees who engage in the corrugated operations of International Paper were disgruntled and asking for pay equity during the COVID19 pandemic. Other replies expressed thanks to International Paper for reducing the waste in their corrugated operations. Moreover, the drill down showed a reply asking International Papers to withdraw their operations from Russia in solidarity with Ukraine. Hence, these faint signals are valuable at 3 strategic levels: First, this sounds the alarm at the human resources level to revise their compensation strategy in order to ensure they are competitive and to avoid operations shutdown due to manpower shortage. Second, this realigns the company's overall strategy to minimize waste and contribute to the circular economy. And third, figure 116 above tells the company that the stakeholders are against doing business in Russia as an act of solidarity with the people of Ukraine. Hence, a call to potentially revise the global value chain strategy or even prepare exit strategies from the Russian market. If not for our data science approach to exploration, these subtle messages could be easily missed due to data overload leading to adverse effects on the competitiveness of the company in the pulp and paper industry.

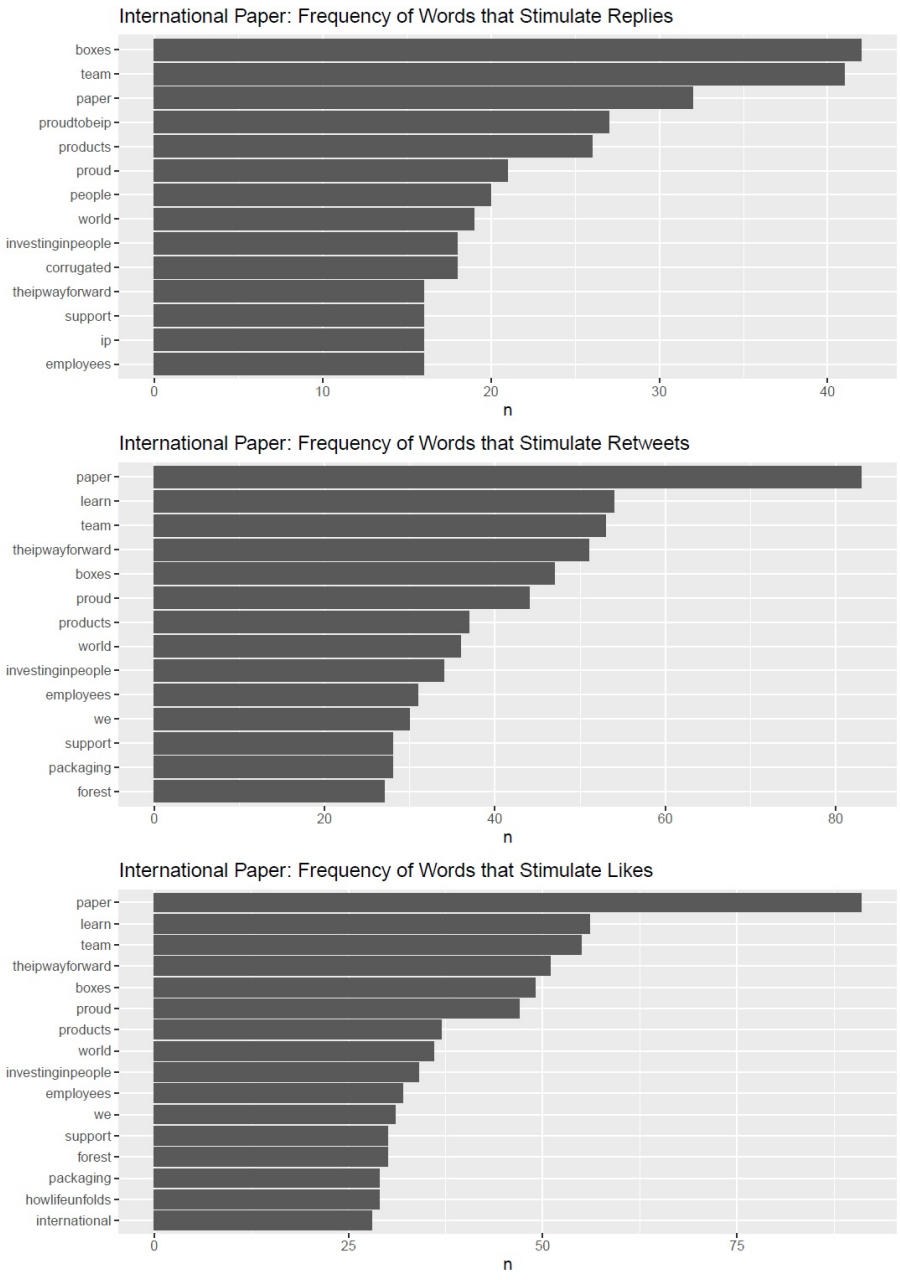


Figure 116: Frequency of words in International Paper's tweets on hot topics

6.1.4 Usage of Most Common Words Over Time

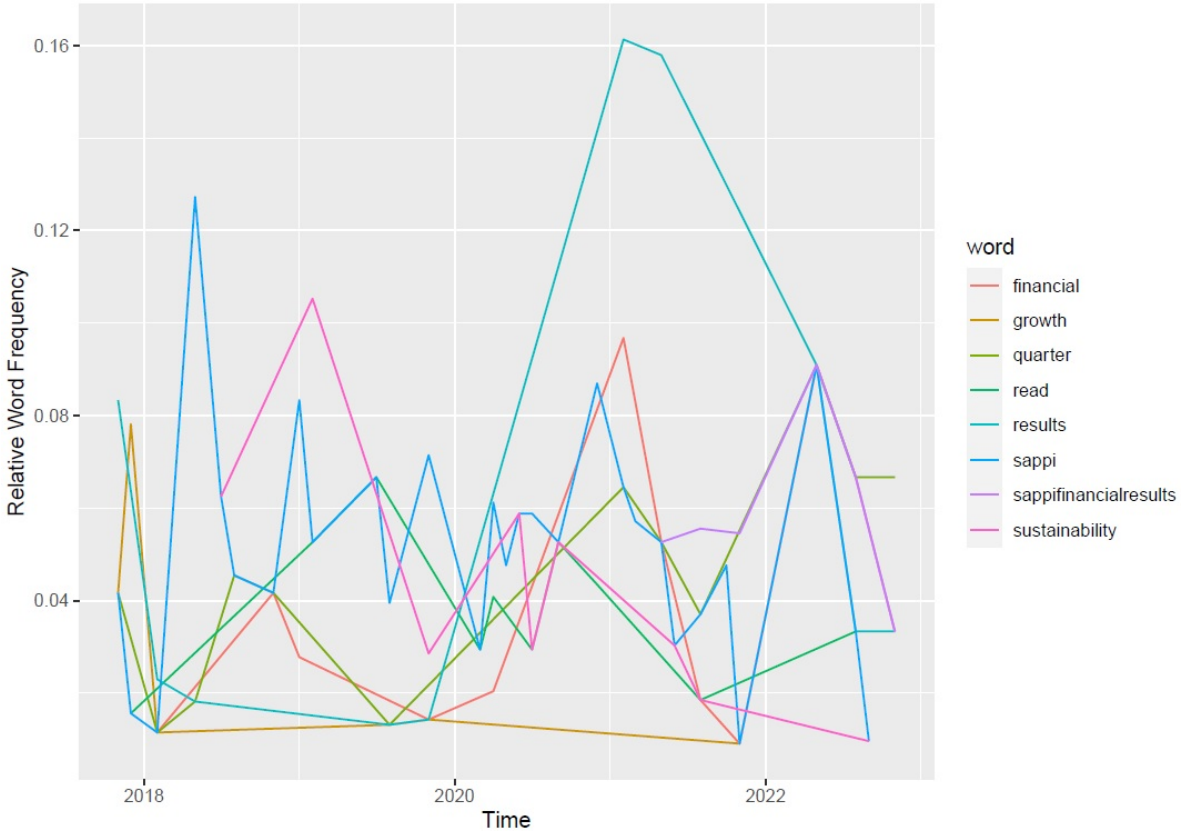


Figure 117: The figure above shows the usage of some of the most common words over integrated time intervals relative to the usage of all other words durnig the same integrated time intervals for Sappi Group

Thanks to the exploration process that is powered by data science, the figure above enables us to study the emergence and evolution of dominant topics discussed on Twitter over time by looking at the number of times a most common word was mentioned in tweets during a time period, divided by the total number of words mentioned in the tweets of the same time period. Yet another measure that is very challenging and resource intensive to evaluate using conventional exploration. In addition, our data science approach produces a standardized and unbiased indication of the evolution of topics when we make comparisons across different twitter accounts since the time period in question depends on the average density of tweets made by the respective twitter account over the whole period of analysis.

In this case, the intermittence and low relative frequency of the word “growth” attracts attention to investigate the strategic expansion moves of a competitor in order to prepare counter strategies knowing that the pulp and paper industry is territorial and price elastic. Moreover, the gradual increase in the relative frequency of the word “results” around the end

of 2019 before showing a spike around the beginning of 2021 provides another faint signal to zoom into. Consequently, the zoom in revealed that Sappi Group was in financial distress and reporting losses in 2020. This subtle signal could be used by Cascades' strategy room in several ways. For example, Cascades could prepare entry strategies to their competitor's market by offering to buy some of the assets they would sell to reverse their losses. Alternatively, Cascades could offer to enter in a strategic alliance, if appropriate, where they could acquire new tacit knowledge from their European partner or access to their intellectual property. In another possible scenario, Cascades could simply learn why Sappi Group experienced losses by performing internal and external analysis in order to avoid the same fate. The aforementioned are just few examples of how our data science powered exploration can inform decision makers and augment their power in the strategy room.

6.1.5 Structural Topic Modelling

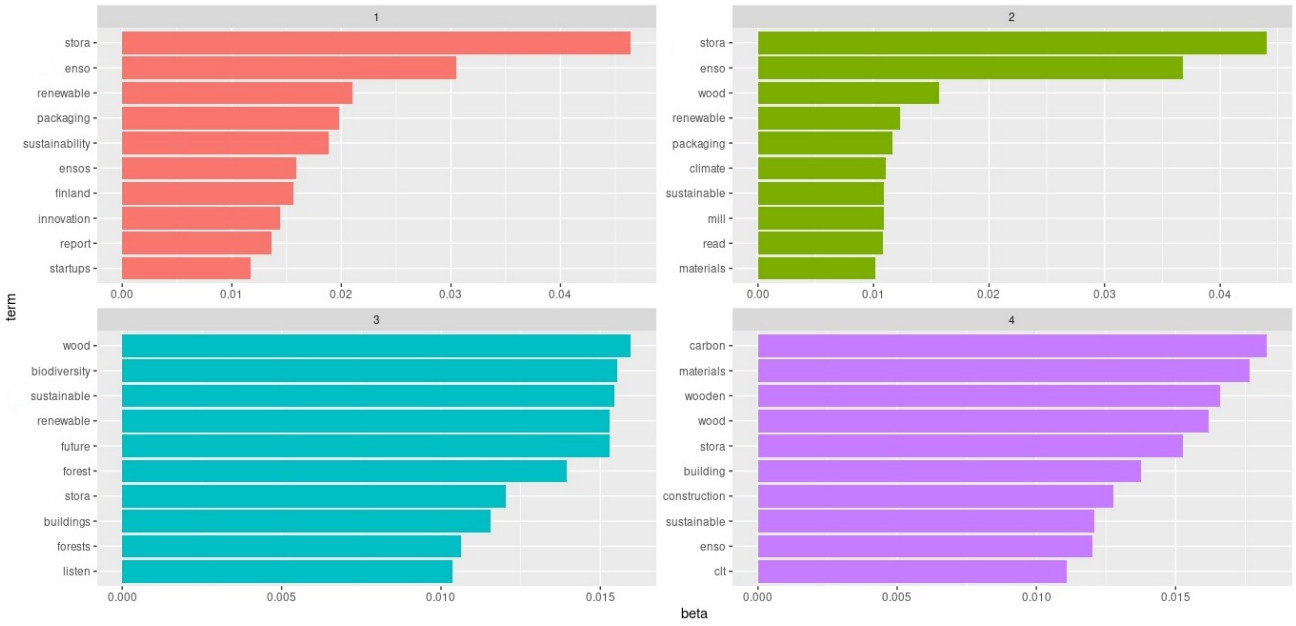


Figure 118: The optimum topic-word combination of the structural topic modelling for Stora Enso's tweets is 10 words and 4 topics

The figure above shows the results of the structural topic modelling for Stora Enso's tweets. The combination of words in topic 1 represents Stora Enso's tweets about how they innovate and their collaborations with startups to produce their products sustainably. As for topic 2, it presents how they contribute towards fighting climate change. Regarding topic 3, Stora Enso's tweets talk about how environment friendly their impact is. Finally, topic 4 encompasses Stora Enso's achievements in the construction industry with their Cross Laminated Timber (CLT).

Thanks to the algorithmic exploration process, NLP revealed that Stora Enso is not just in the pulp, paper and packaging industry. They have also pivoted into the construction industry. The power of data science to summarize large amounts of data and organize them in such a layout empowers decision makers in the strategy room to save time and resources in the exploration process. In other words, this provides Cascades with a starting point to craft a winning strategy that could gain them a competitive advantage in the pulp and paper industry and elsewhere.

6.1.6 Sentiment Analysis

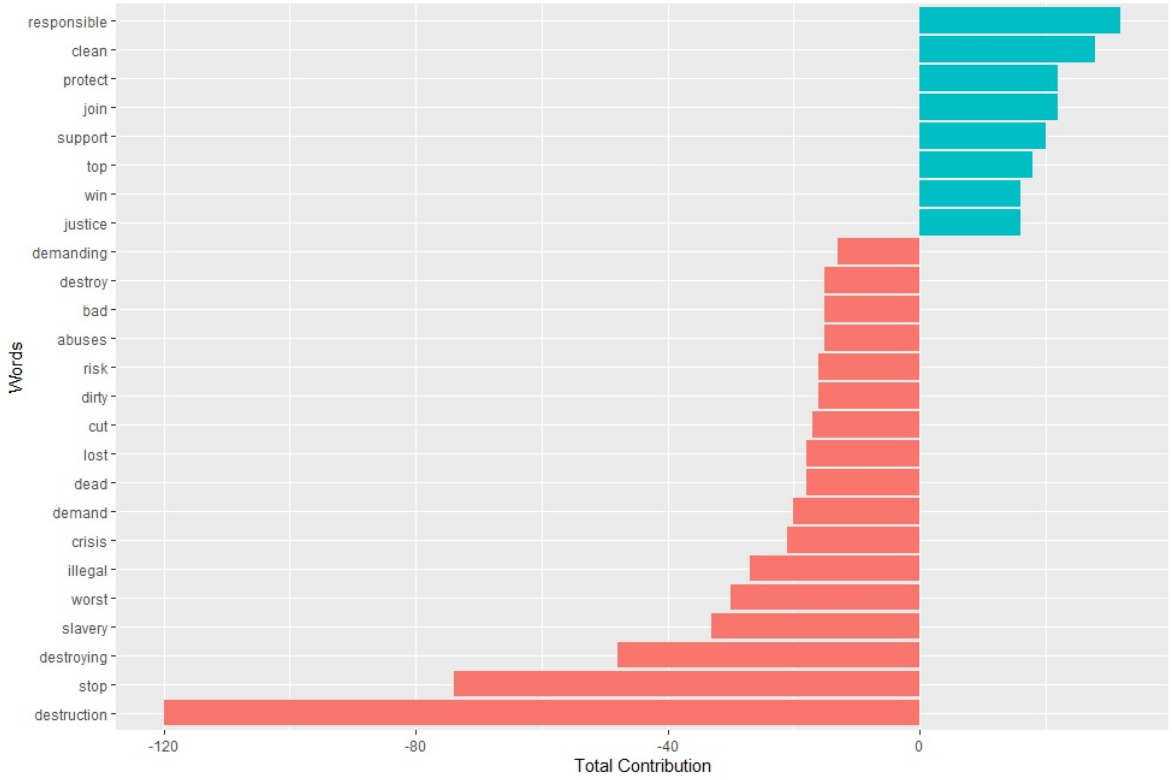


Figure 119: Sentiment analysis for Mighty Earth's qualifying tweets as a whole. The figure shows the words with the highest contribution to the positive as well as the negative sentiment of the tweets. The total contribution is evaluated as the afinn contribution score for each word multiplied by the total number of occurrences.

As a lobby group that has the power to influence stakeholders through criticizing malpractices, the sentiment analysis shown above provides a very important summary of their discussions on Twitter. Just like the previous subsections, this could be used by Cascades strategy room to identify the most negative topics, correlate them with the activity counts to identify what the twitter audience voted for as the worst malpractice before crafting ESG strategies accordingly. But we also find that running a sentiment analysis on the crowds' replies to Mighty Earth's tweets would be very useful for Cascades' decision makers. Especially if the said tweets garner a very large number of replies that would consume lots of time and resources to go through and analyse manually. Hence, it would give a good indication to what malpractices should be corrected or avoided with the limited resources available. If going through the sheer number of replies and analyzing their contents were to be done manually or even through conventional methods such as surveys, the exploration process would be prone to biases in addition to consuming a large amount of resources. This again proves the huge benefits of integrating data science into the exploration process of organisational ambidexterity.

6.1.7 Overall Network of Words in Tweets

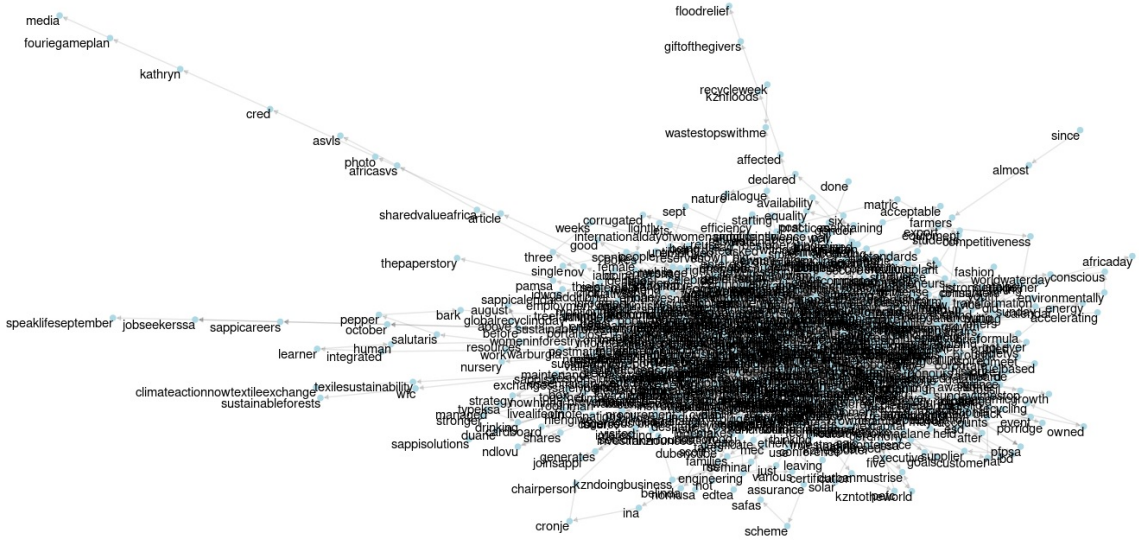


Figure 120: This directed network graph is showing the overall communication strategy of Sappi Southern Africa on Twitter. The labelled nodes represent all the words used and the directed edges indicate how Sappi Southern Africa structures their phrases in their different messages from the beginning to the end. As the darkness of the edges increases, the frequency of the usage of the second word is higher. Naturally, the graph crowdedness would increase as the number of tweets analysed increases. But using a larger screen with a higher resolution would reveal the words with highest betweenness centrality shown in the middle of the whole network

The network graph shown above maps the communication strategy of Sappi Southern Africa. This provides a useful guide for Cascades' decision makers to structure competitive tweets that appeal to the twitter crowd while also converging to their values and mission similar to what is shown in the network above. A good example of that is Sappi Southern Africa's sustainable practices that start to converge from the top to the dense part of the network. Another example of that is their focused PR efforts that start to converge from the top left to the dense part of the network. An additional example is their talent attraction effort that starts converging from the left side of the dense network. The convergence of all the different topics to one of few clusters of words within the dense area of the network shows how all of Sappi Southern Africa's practices tie back to their values or mission.

Using conventional means, the exploration process would require going manually through numerous news releases, electronic articles, annual company generated reports and other sources of information continuously to analyze and summarize what the competitor is all about. For that reason, decision makers in the strategy room shy away from directing resources into the exploration process as the expected return on investment into the

exploration process is not guaranteed. But what we did here, thanks to data science, is facilitate the exploration process to supply streamlined sets of information to the finger tips of decision makers in the strategy room.

6.1.8 Network of Most Common Used Pairs of Words

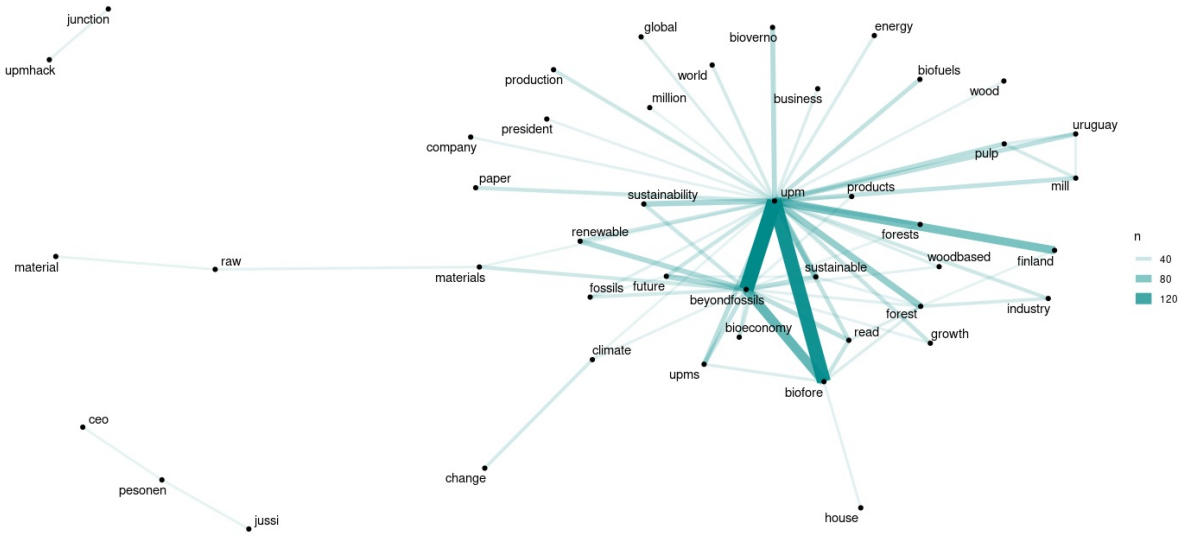


Figure 121: This network graph is showing the most common usage of the pair of words tweeted by UPM Global. The labelled nodes represent the words and the edges connect the pair of words used together. As the thickness and darkness of the edges increases, the frequency of appearance of both words together in the tweet increases.

The map above is a stark evidence on how network analysis could help in identifying pivot opportunities by just analyzing the 3 stakeholders we referred to in our radar detection radius. The interconnections between the different small worlds demonstrates the diverse and innovative business models that UPM Global implements, including their pivot from the pulp and paper industry to the circular economy and the energy sector. It is also interesting to note how the algorithm generated graph recognized that the CEO of UPM Global is Jussi Pesonen. The graph also shows how the beyondfossils strategy is almost as central as UPM Global itself to the other topics within the graph since both words are connected to almost the same words. This indicates that the beyondfossils strategy is an integral part of the UPM brand.

Besides that, we strongly believe that the network analysis of entities beyond our detection radius, i.e. outside the pulp and paper industry, could unleash the real powers of network analysis to uncover pivot opportunities. This was not attempted during our dissertation because it is a topic that is too advanced for the scope of our proof of concept and could form a standalone subject for a dissertation. Nevertheless, a supporting proof of its potential

success is how NASA got inspired by lobsters to create the James Webb Telescope (BBC World Service 2021).

It is important to note that all of the steps shown above could either be used sequentially or separately to uncover insights in tweets. In some cases, the results from one of these steps is enough to paint a clear picture of the competitive landscape. While in other cases, more analysis and correlation with other steps is necessary to paint that clear picture for decision makers to move forward with more confidence.

6.2 Limitations

While integrating data science into the exploration process proved its huge potential to augment an organization's strategic decision making, it still has limitations like any other human generated product. To being with, this analysis could not be carried out if the stakeholders are not on Online Social Networks (OSNs) in the first place. To circumvent that, our data science methodology could be used instead to analyze electronic publications on competitors and shareholders if readily available. Though the disadvantages related to that approach are as follows: First, it would not work on private organisations who do not publish any information about their yearly operations or results. Second, a dedicated team needs to be subscribed to follow up continuously with the electronic publications which is not practical and would be similar to the conventional exploration process. And most importantly, it could create lags in decision making between the time of the publication until the analysis of the data, such as analyzing annual reports of organisations that are published only once per year which brings us to our next limitation.

In addition, it would be interesting to envision being in the strategy room, equipped with a real-time global perspective on public discourse, with updates streaming in synchrony with every breath. Such immediacy offers decision-makers a live pulse of global conversations, enabling them to act based on the most recent data. This agility in decision-making not only ensures relevance but also provides a significant competitive advantage, as underscored in our literature review. In the dynamic landscape of business, decisions anchored in current, non-stale data are paramount for success. However, the use of the rest API in our methodology does present a limitation in this regard, as it primarily facilitates the analysis of historical data. While we employed the rest API to demonstrate our proof of concept, the streaming API holds greater promise for continuous, real-time monitoring of Twitter conversations. This capability ensures that strategic decisions are informed by the freshest

insights, emphasizing the importance of our subsequent discussions.

It is important to highlight that our data wrangling, especially our filtering mechanism, is not a one size fits all and it could take different shapes depending on the type of data, the quantity of data, the type of industry, industry dynamics... etc. So how do we decide what is the elimination threshold that would tell us whether a tweet is important or not? It really goes by the saying, if everything is important then nothing is important. Again, this is subjective and could change depending on how we define importance. In our point of view, we picked the 3^{rd} quartile of activity total because we want to look at conversations that were voted most important by the crowd. This is not biased because it is the crowd who decided and not us. In other words, if the disqualified tweets with activity total less than the 3^{rd} quartile were deemed important by the crowd, they would have garnered more attention from the crowd which gets translated into more activity putting them potentially above the 3^{rd} quartile. This helps us eliminate the misleading noise with our sole dependence on ternary ratio spikes. Even though our filtering mechanism and the elimination thresholds we set may still be seen as too constrictive by some or too loose by others, we justified here the reason behind them and proved their success with our results.

It is also important to highlight that we were not prone to survivorship bias by picking companies based on highest turnover value, since we are looking at the real competitiveness level which could be measured by turnover as an example, i.e companies cannot jump several ranks at once. Nonetheless, this exercise could still be done for larger number of twitter accounts simultaneously in real time.

Another potential limitation is present with the use of graphs to visually identify spikes in hot topics or spikes in ternary ratios since this could be prone to errors in identifying the exact time intervals for analysis. While this could simply be resolved by identifying % change in activity total between different time intervals of the whole period of analysis, it would generate further complications that require us to set a threshold of how much of a % change deems this tweet to be considered "spiky". Regardless, we succeeded in providing our proof of concept using the visual graphs.

Moreover, we find that the sentiment analysis is subjective by default. The reason behind that is the AFINN lexicon is generated by humans and the determination of scores for each word in the lexicon may differ from one person to another. Regardless, it would not have a sizable effect on the overall results of the sentiment analyses we ran since the change in perception of

how positive or negative a word is will not be different to the extent that it would flip the overall contribution. In addition, the lexicons do not contain all the productive words in the English dictionary. Furthermore, the lexicons are built for sentiment analysis in English and thus, this methodology would not work on tweets made in other languages unless they are translated (Nielsen 2011).

Add to that, the limitation of the data collected is that it does not show the actual replies to the tweets made. We strongly believe that running our analysis on the replies of the stakeholders in addition to the hot tweets could help paint a clearer picture of the two-way conversations happening on OSNs.

6.3 Summary

To sum up, our research endeavors have in fact detailed conclusive responses to our posed questions. We have successfully demonstrated a proof of concept, underscoring the transformative potential of data science in managing the flood of data and refining strategic decision-making. Thus, cracking the paradox of administration (Thompson 2003).

Moreover, our proof of concept has the ability to provide the pertinent data we need in real-time using less resources. This perspective consequently increases an organisation's flexibility, (Abernathy 1978; Benner and Tushman 2003), minimizes the risk of being rendered irrelevant in a rapidly changing and competitive environment (March 1991; Lis et al. 2018) and diminishes the likelihood of an organisation's failure in the face of change (March 1991; Michael L. Tushman and O'Reilly 2013). Subsequently, performing exploration using our data science approach enables organisations to reallocate the scarce resources from exploration to exploitation. Thus, allowing organisations to simultaneously explore and exploit, i.e. be ambidextrous, and strengthening their ability to compete over time (Benner and Tushman 2015; P. Ghemawat and Ricart Costa 1993; March 1991; Michael L. Tushman and O'Reilly 2013; Weick 1979; Teece, Pisano, and Shuen 1997; C. A. O'Reilly and Tushman 2011). Furthermore, this contributes to reducing the biases away from exploration due to the perception of uncertainty associated with it (Benner and Tushman 2003; March 1991).

As for the strategic advantages, our proof of concept provides a birds-eye view of the strengths, weaknesses, threats and opportunities within the competitive landscape using less resources. These advantages were demonstrated throughout all the text mining steps of our analysis of the tweets. In addition, the ability to search for and discover, i.e. explore, the

latest trends and the changing tastes of the stakeholders enables organisations to integrate the findings from NLP into their long-term growth strategies that may include market entries or even industry penetrations. In other words, pivot opportunities. Equally important, the data science-powered network analysis proved its ability to uncover nuances in competitors' business models in addition to their communication strategies.

While our proof of concept remains foundational, its adaptability is evident, making it applicable across diverse industries, purposes, sectors, and on various OSNs. At its most rudimentary level, data science offers a lens to discern, sift through, and structure vast datasets, thereby illuminating the path for informed decisions. In a more advanced scenario, it can unveil insights that might have otherwise remained obscured. Yet, it's crucial to recognize that the realm of data science is still in its nascent stages. And given our current understanding, it's challenging to delineate the full extent of its potential in enhancing our decision-making processes. The horizon of possibilities remains vast and largely uncharted.

Now, does our proof of concept imply that the end is near for human decision making? Definitely not, data science in addition to the algorithmic models built upon it will not and should not take the place of humans. Instead, it will augment their abilities. We proved that by having numerous ears and eyes in different places around the world simultaneously. Most importantly, data science allows us to do that on the whole population instead of just a sample. Hence, augmenting the decision making abilities of the strategy room even at the macro level. What remains is under the control of one brain, our brain, to make the decisions based on the resulting info. That would be step 4 of figure 1 in the introduction section.

So whenever I am challenged by such dilemmas, I always imagine myself sitting with Alfred Nobel, Hassan Kamel Sabbah, Charles Kettering, Thomas Edisson, Nikola Tesla or any other noteworthy inventor to ask them if they regret their ingenuity. Because if it wasn't for Nobel, it would be difficult to visualize how metals could have been extracted to pave the way for the industrial revolution ("Alfred Nobel's Life and Work" 2022). And if it wasn't for Charles Kettering ("Kettering Aerial Torpedo 'Bug'" 2015), we potentially wouldn't be able to use drones to monitor and improve our agricultural lands (Royer et al. 2021). And if it wasn't for all the prominent scholars I cited in this dissertation, I wouldn't have been able to build upon their remarkable contributions to present this thesis.

Regarding future research avenues, they are as vast as the data science field. But what I imagine seeing is the use of data science to augment decision making in the public sector, such

as integrating data science into Michael Porter's Diamond (M. E. Porter 1990). Another interesting area of research would be to build upon the graph theory knowledge using data science, to uncover pivot opportunities by analyzing commonalities between different unrelated industries.

As a final remark, Sir Francis Bacon is still right after 400+ years - knowledge is power indeed (Bacon 1620). Friedrich Hayek on the other hand might have been a bit short-sighted by thinking that everyone may only know about their little corner of the economy but much less elsewhere (Hayek 1945). As a matter of fact, we proved that one single entity could possess complete knowledge of everything available in the present using data science. After that, it is up to us to make the right decisions.

Bibliography

Abernathy, William J. 1978. *The Productivity Dilemma : Roadblock to Innovation in the Automobile Industry*. Johns Hopkins University Press.

Adner, Ron. 2002. "When Are Technologies Disruptive a Demand-Based View of the Emergence of Competition." Strategic Management Journal.

Afuah, Allan, and Christopher L. Tucci. 2013. "Value Capture and Crowdsourcing." *AMR* 38 (3): 457–60. https://doi.org/10.5465/amr.2012.0423.

Albert, Reka, and Albert-Laszlo Barabasi. 2002. "Statistical Mechanics of Complex Networks." *Rev. Mod. Phys.* 74 (1): 47–97. https://doi.org/10.1103/RevModPhys.74.47.

"Alfred Nobel's Life and Work." 2022. *NobelPrize.org*. https://www.nobelprize.org/alfred-nobel/alfred-nobels-life-and-work/.

"Ambidextrous Definition & Meaning - Merriam-Webster." n.d. https://www.merriam-webster.com/dictionary/ambidextrous. Accessed January 18, 2023.

Ante, Lennart. 2023. "How Elon Musk's Twitter Activity Moves Cryptocurrency Markets." *Technological Forecasting and Social Change* 186 (January): 122112. https://doi.org/10.1016/j.techfore.2022.122112.

Anthony, Scott D. 2016. "Kodak's Downfall Wasn't About Technology." *Harvard Business Review*, July.

Bacon, Sir Francis. 1620. *Novum Organum*. P. F. Collier.

Barabási, Albert-László. 2003. "NETWORK SCIENCE RANDOM NETWORKS."

BBC World Service. 2021. "The Space Telescopes Inspired by Lobsters - BBC World Service."

Beckmann, S. C., M. Morsing, L. Reisch, M Morsing, and S. C Beckmann. 2006. *Strategic CSR Communication*. Edited by null. Vol. null. Null.

Benner, Mary J. 2009. "Dynamic or Static Capabilities? Process Management Practices and Response to Technological Change." *Journal of Product Innovation Management* 26 (5): 473–86. https://doi.org/10.1111/j.1540-5885.2009.00675.x.

Benner, Mary J., and Michael L. Tushman. 2003. "Exploitation, Exploration, and Process Management." *The Academy of Management Review* 28 (2): 238. https://doi.org/10.2307/30040711.

———. 2015. "Reflections on the 2013 Decade Award: 'Exploitation, Exploration, and Process Management: The Productivity Dilemma Revisited' Ten Years Later." *AMR* 40 (4): 497–514. https://doi.org/10.5465/amr.2015.0042.

Blei, David M. 2003. "Latent Dirichlet Allocation."

Brown, Shona L., and Kathleen M. Eisenhardt. 1998. *Competing on the Edge : Strategy as Structured Chaos*. Harvard Business School Press.

Chesbrough, Henry W. 2006. *Open Business Models : How to Thrive in the New Innovation Landscape*. Harvard Business School Press.

Chizobah, Morah. 2022. "How Companies Calculate Revenue." *Investopedia*. https://www.investopedia.com/ask/answers/09/how-companies-calculate-revenue.asp.

CIRANO. 2023. "About Us - CIRANO." https://cirano.qc.ca/en/about/cirano.

Cohen, Wesley M, and Daniel A. Levinthal. 1990. "Absorptive Capacity: A New Perspective on Learning and Innovation."

Cornall, Jim. 2022. "Elopak and Nippon Paper Agree Oceania License." *Elopak and Nippon Paper Agree Oceania License*. https://www.dairyreporter.com/Article/2022/04/05/elopak-and-nippon-paper-agree-oceania-license.

Davis, Doug. 2011. "Play-Doh Forms Memorable Career." *The Daily News Journal*, n/a.

Desjardins, Jeff. 2019. "How Much Data Is Generated Each Day? | World Economic Forum." https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/.

Eaves, Elisabeth. 2006. "Free Money." *Foreign Policy*, no. 154: 92.

Ebben, Jay J., and Alec C. Johnson. 2005. "Efficiency, Flexibility, or Both? Evidence Linking Strategy to Performance in Small Firms." *Strat. Mgmt. J.* 26 (13): 1249–59. https://doi.org/10.1002/smj.503.

FUJIFILM, FUJIFILM. 2007. "ASTALIFT." https://www.fujifilm.com/sg/en/consumer/skincare/astalift.

Fujii, Hide. 2016. "Fujifilm: Surviving the Digital Revolution in Photography Through Diversification into Cosmetics - Technology and Operations Management." https://d3.harvard.edu/platform-rctom/submission/fujifilm-surviving-the-digital-revolution-in-photography-through-diversification-into-cosmetics/.

Geerts, Annalies, Floortje Blindenbach-Driessen, and Paul Gemmel. 2010. "Achieving a Balance Between Exploration and Exploitation in Service Firms: A Longitudinal Study." *Paper Presented at the Annual Meetings of the Academy of Management*.

Ghemawat, Pankaj. 2007. *Redefining Global Strategy : Crossing Borders in a World Where Differences Still Matter*. Boston, Massachusetts: Harvard Business School Press.

Ghemawat, Pankaj, and Joan E. I Ricart Costa. 1993. "The Organizational Tension Between Static and Dynamic Efficiency." *Strat. Mgmt. J.* 14 (S2): 59–73.

https://doi.org/10.1002/smj.4250141007.

Glaveski, Steve. 2021. "The Top 10 Company Pivots of All-Time." *Steve Glaveski*.

Goyal, Sanjeev, Marco J. Van Der Leij, and José Luis Moraga-González. 2006. "Economics: An Emerging Small World." *Journal of Political Economy* 114 (2): 403–12. https://doi.org/10.1086/500990.

Harding, Robin. 2010. "Twitter Faces a Battle for Tweets." https://www.ft.com/content/8a0009f2-2df2-11df-b85c-00144feabdc0.

Hayek, F A. 1945. "THE USE OF KNOWLEDGE IN SOCIETY."

Huang, Wayne, John Mitchell, Carmel Dibner, Andrea Ruttenberg, and Audrey Tripp. 2018. "How Customer Service Can Turn Angry Customers into Loyal Ones." *Harvard Business Review*, January.

IBM. 2020. "What Is Natural Language Processing? | IBM." https://www.ibm.com/topics/natural-language-processing.

———. n.d. "What Is Data Science? | IBM." https://www.ibm.com/topics/data-science. Accessed June 25, 2023.

Karikallio, Hanna, Petri Mäki-Fränti, and Niko Suhonen. 2011. "Competition in the Global Pulp and Paper Industries – An Evaluation Based on Three Approaches." *Journal of Forest Economics* 17 (1): 91–104. https://doi.org/10.1016/j.jfe.2010.09.004.

"Kettering Aerial Torpedo 'Bug'." 2015. *National Museum of the United States Air Force™*. https://www.nationalmuseum.af.mil/Visit/Museum-Exhibits/Fact-Sheets/Display/Article/198095/kettering-aerial-torpedo-bug/https%3A%2F%2Fwww.nationalmuseum.af.mil%2FVisit%2FMuseum-Exhibits%2FFact-Sheets%2FDisplay%2FArticle%2F198095%2Fkettering-aerial-torpedo-bug%2F.

Klara, Robert. 2016. "Play-Doh." *Adweek* 57 (30): 27–28.

Kouloukoui, Daniel, Nathalie de Marcellis-Warin, Sonia Maria da Silva Gomes, and Thierry Warin. 2023. "Mapping Global Conversations on Twitter about Environmental, Social, and Governance Topics Through Natural Language Processing." *Journal of Cleaner Production* 414 (August): 137369. https://doi.org/10.1016/j.jclepro.2023.137369.

Kuhlberg, M. 2015. "Pulp and Paper Industry. In the Canadian Encyclopedia." https://www.thecanadianencyclopedia.ca/en/article/pulp-and-paper-industry.

Lacka, Ewelina, D. Eric Boyd, Gbenga Ibikunle, and P. K. Kannan. 2022. "Measuring the Real-Time Stock Market Impact of Firm-Generated Content." *Journal of Marketing* 86

(5): 58–78. https://doi.org/10.1177/00222429211042848.

Levinthal, Daniel A. 1991. "Organizational Adaptation and Environmental Selection-Interrelated Processes of Change." *Organization Science* 2 (1): 140–45. https://doi.org/10.1287/orsc.2.1.140.

Levinthal, Daniel A., and James G. March. 1993. "The Myopia of Learning." *Strat. Mgmt. J.* 14 (S2): 95–112. https://doi.org/10.1002/smj.4250141009.

Lis, Andrzej, Barbara Józefowicz, Mateusz Tomanek, and Patrycja Gulak-Lipka. 2018. "The Concept of the Ambidextrous Organization: Systematic Literature Review." *IJCM* 17 (1). https://doi.org/10.4467/24498939IJCM.18.005.8384.

Lundmark, Leif W., Chong Oh, and J. Cameron Verhaal. 2017. "A Little Birdie Told Me: Social Media, Organizational Legitimacy, and Underpricing in Initial Public Offerings." *Information Systems Frontiers* 19 (6): 1407–22. https://doi.org/10.1007/s10796-016-9654-x.

Luo, Yadong, and Huaichuan Rui. 2009. "An Ambidexterity Perspective Toward Multinational Enterprises From Emerging Economies." *Academy of Management Perspectives*.

Maeir, Jens. 2015. *The Ambidextrous Organization : Exploring the New While Exploiting the Now*. Houndmills, Basingstoke Hampshire ; New York, NY : Palgrave Macmillan.

March, James G. 1991. "Exploration and Exploitation in Organizational Learning" 2 (1): 71–87.

Minot, Joshua. 2022. "Gauge Against the Machine: Improving Representations Within Sociotechnical Instruments to Enrich Context and Identify Biases."

Mohammad, Saif M., and Peter D. Turney. 2011. "NRC Emotion Lexicon."

Morsing, M, and S. C Beckmann. 2006. *Strategic CSR Communication*. Edited by null. Vol. null. Null.

Morsing, Mette, Majken Schultz, and Kasper Ulf Nielsen. 2008. "The 'Catch 22' of Communicating CSR: Findings from a Danish Study." *Journal of Marketing Communications* 14 (2): 97–111. https://doi.org/10.1080/13527260701856608.

Nelson, Richard R., and Sidney G. Winter. 1982. *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press.

"Netflix Reed Hastings." 2017. *Game Changers*. [Place of publication not identified]: Bloomberg.

Nielsen, F. {\AA}. 2011. "AFINN."

http://www2.compute.dtu.dk/pubdb/pubs/6010-full.html.

O'Reilly, Charles A. III, and Michael L. Tushman. 2004. "The Ambidextrous Organization." https://hbr.org/2004/04/the-ambidextrous-organization.

O'Reilly, Charles A., and Michael L. Tushman. 2011. "Organizational Ambidexterity in Action: How Managers Explore and Exploit." *California Management Review* 53 (4): 5–22. https://doi.org/10.1525/cmr.2011.53.4.5.

Porter, Michael E. 1990. "The Competitive Advantage of Nations." Free Press.

Porter, Scott. 2021. "How Twitter Has Become a Key Customer Support Channel." https://business.twitter.com/en/blog/how-twitter-has-become-a-key-customer-support-channel.html.

Qiu, Guang, Bing Liu, Jiajun Bu, and Chun Chen. 2011. "Opinion Word Expansion and Target Extraction Through Double Propagation." *Computational Linguistics* 37 (1): 9–27. https://doi.org/10.1162/coli_a_00034.

Royer, Annie, Nathalie de Marcellis-Warin, Ingrid Peignier, and Thierry Warin. 2021. "LA RÉVOLUTION NUMÉRIQUE APPLIQUÉE À L'AGRICULTURE AU QUÉBEC."

Sutton, Robert. 2002. "Look at It Another Way You Can't See a Creative Solution... You're Playing Safe... A Victim of Your Own Expertise. Try a Little Reverse Deja-Vu." *Financial Times*, 12.

Teece, David J., Gary Pisano, and Amy Shuen. 1997. "Dynamic Capabilities and Strategic Management." *Strat. Mgmt. J.* 18 (7): 509–33. https://doi.org/10.1002/(SICI)1097-0266(199708)18:7%3C509::AID-SMJ882%3E3.0.CO;2-Z.

Thompson, James D. 2003. *Organizations in Action: Social Science Bases of Administrative Theory.* Classics in Organization and Management. New Brunswick, NJ: Transaction Publishers.

Tushman, Michael L., and III Charles A. O'Reilly. 1996. "Ambidextrous Organizations: Managing Evolutionary and Revolutionary Change." *California Management Review* 38 (4): 8–29. https://doi.org/10.2307/41165852.

Tushman, Michael L., and Charles A. O'Reilly. 1999. "Building Ambidextrous Organizations: Forming Your Own "Skunk Works"." *Healthcare Forum Journal* Volume 42, Issue 2; ISSN: 0899-9287 (March): 20.

Tushman, Michael L, and Charles A. III O'Reilly. 2013. "Organizational Ambidexterity: Past, Present and Future." *Academy of Management Perspectives (in Press).*

Twitter. 2022. "About Twitter Circle." *About Twitter Circle.*

https://help.twitter.com/en/using-twitter/twitter-circle.

———. 2023a. "Stream Tweets in Real-Time." *Stream Tweets in Real-Time*. https://developer.twitter.com/en/docs/tutorials/stream-tweets-in-real-time.

———. 2023b. "Twitter API Documentation." *Twitter API Documentation*. https://developer.twitter.com/en/docs/twitter-api.

Twitter, Twitter. 2014. "The 2014 #YearOnTwitter." https://blog.twitter.com/en_us/a/2014/the-2014-yearontwitter.

———. 2017. "How Twitter Transforms Conversations Between Companies and Customers." https://marketing.twitter.com/en/perspectives/twitter-transforms-conversations-between-companies-and-customers.

———. 2020. "Twitter Privacy Policy." https://twitter.com/en/privacy/previous/version_16.

"User-Generated Internet Content Per Minute 2022." 2023. *Statista*. https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/.

Uzzi, Brian, and Jarrett Spiro. 2005. "Collaboration and Creativity: The Small World Problem." *American Journal of Sociology* 111 (2): 447–504. https://doi.org/10.1086/432782.

Warin, Thierry. 2022. "The World Health Organization in a Post-COVID-19 Era: An Exploration of Public Engagement on Twitter." CIRANO. https://doi.org/10.54932/EHUH4224.

Warin, Thierry, and CIRANO. 2020. "MATH60033A: [Course] Quantitative Methods in International Business Research." *MATH60033A*. https://warin.ca/math60033a/syllabus.html#tldr.

Watts, Duncan J, and Steven H Strogatz. 1998. "Collective Dynamics of 'Small-World' Networks" 393.

Weick, Karl E. 1979. *The Social Psychology of Organizing*. 2nd ed. Topics in Social Psychology. Addison-Wesley.

"Western Union Stops Telegrams." 2006. *The Guelph Mercury*, A7.

Zhang, Lei, Shuai Wang, and Bing Liu. 2018. "Deep Learning for Sentiment Analysis: A Survey." *WIREs Data Mining Knowl Discov* 8 (4). https://doi.org/10.1002/widm.1253.