Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

> *"It's a good sign that this sector isn't niche anymore. It's actually attracting some of the most prolific brand-name funds because of the potential. … It's not just funding that has accelerated, I genuinely feel that the ecosystem shifted in such a way to allow for more rapid and scaled adoption."*
>
> Megan Zweig
> Rock Health's Chief Operating Officer

Categorizing the market focus of larger samples of companies can be a tedious and time-consuming process for both researchers and business analysts interested in developing insights about emerging business sectors. The objective of this article is to suggest a text analytics approach to categorizing the application areas of companies operating in the digital health sector based on the information provided on their websites. More specifically, we apply topic modeling on a collection of text documents, including information collected from the websites of a sample of 100 innovative digital health companies. The topic model helps in grouping the companies offering similar types of market offers. It enables identifying the companies that are most highly associated with each of the topics. In addition, it allows identifying some of the emerging themes that are discussed online by the companies, as well as their specific market offers. The results will be of interest to aspiring technology entrepreneurs, organizations supporting new ventures, and business accelerators interested to enhance their services to new venture clients. The development, operationalization, and automation of the company categorization process based on publicly available information is a methodological contribution that opens the opportunity for future applications in research and business practice.

1. Introduction

To identify new business ideas and shape opportunities, the founders of new ventures should systematically scan the market environment and search for new information, associate previously unconnected information, and make judgments about the commercialization potential of their business ideas (Tang et al., 2012). This is especially important in emerging technology sectors, such as digital health, where the increased need for novel digital health solutions has inspired many startups to branch into it. One of the ways for new ventures to address this business intelligence need is to benefit from the increasingly available open source web search and text analytics tools as a way to search for, collect, and analyze publicly available competitor information about newly introduced market offers and innovative applications of new and emerging technologies. This can help them generate valuable insights about their competitive market offer (Tanev et al., 2015).

Developing business analytics capabilities offers a research opportunity for scholars who can focus on suggesting and validating applied analytical methods that help new companies interested to engage in market intelligence activities at the early stages of their existence. Developing such capabilities can be highly valuable for researchers themselves, giving them an opportunity to identify innovation trends related to the early stages of commercializing new and emerging technologies.

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

The objective of this article is to apply topic modeling (Blei et al., 2012; Hecking & Leydesdorff, 2018) on collections of text documents that include information from the websites of 100 innovative digital health companies and categorize them in terms of their overall business application areas and market offer types. The list of companies was created by the Medical Futurist Organization (https://medicalfuturist.com/the-top-100-digital-health-companies-an-infographic) through the evaluation of four selection criteria: innovation mindset, disruptiveness of technology, business model viability, and clear dedication to digital health domain. The study is part of the deliverables of a research program that was designed to examine the applicability of the topic modeling approach to mapping technology companies in terms of their market offer types (Johnson et al., 2008) and, eventually, to develop business opportunity maps that could be used by new ventures interested to enter a specific business domain (Mamosian et al., 2018).

The research question is: How to apply topic modeling on textual information provided on the websites of a sample of digital health companies to identify their business application areas, dominant market offer types, and most representative companies engaged in offering them? In this sense, the article aims to make a contribution to the literature by shaping an automated text analytics approach to the categorization of larger samples of companies in a specific business domain. One of the key benefits of the suggested approach for technology innovation and entrepreneurship researchers is helping to easily identify highly representative cases within larger samples of companies. Such identification allows for methodological enrichment through the valuable integration of quantitative and qualitative research inquiries related to emerging business sectors and the commercialization of new technologies.

The content of the article is organized as follows. The next section summarizes insights from literature on the innovation and entrepreneurial potential of the digital health sector, as well as introducing the topic modeling approach as an example of a Natural Language Processing (NLP) technique that could be highly valuable for developing of competitive insights for early-stage companies. The third section describes the methodology adopted in our study. The next section includes a summary of the topic modeling results and key findings emerging from post-processing analysis. The article ends with a conclusion describing the contributions and the potential of the method for future studies.

2. Literature review

2.1 The entrepreneurial potential of the digital health sector

Healthcare is entering a new digital era where telemedicine, virtual reality, robotics, smart phones, and other technological advancements are slowly becoming part of regular healthcare practices (Wulfovich & Meyers, 2020). Digital health technologies offer a way to change many of the current challenges faced by healthcare systems. They have the potential to transform the medical field by improving patient outcomes, increasing quality of healthcare, and reducing costs.

A recent study by the Medical Futurist Organization indicates that the global digital health market is expected to exceed $504.4 billion USD by 2025, which is a nearly six-fold increase from its $86.4 billion USD 2018 valuation. The COVID-19 pandemic has accelerated the adoption of digital health and will continue fostering its growth in the near future. For example, one of the growing niches of the digital health market, which is attracting the attention of many aspiring technology entrepreneurs, includes startups that use artificial intelligence in drug discovery, disease diagnosis, and patient monitoring.

Digital health entrepreneurs pursue opportunities under conditions of high uncertainty. They focus on creating healthcare stakeholder value through the deployment of digital health innovations. However, the emergence of the business sector is associated with significant challenges in identifying opportunities, along with the design, development, testing, and commercialization of digital health technologies and products. The challenges of digital health entrepreneurship and innovation start at the very beginning of the technology commercialization roadmap, with a) industry and market analysis, and b) opportunity identification and assessment (Wulfovich & Meyers, 2020). It is important therefore for the founders of new digital health companies to develop the analytical skills they need to carefully examine the latest market offers and value propositions in their business sector, in order to be able

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

make differentiations in the marketplace.

2.2 Applications of topic modeling to the categorization of the market offers of innovative firms

Our study uses topic modeling to examine the corpus of text documents corresponding to the webpages of innovative digital health companies. Topic modeling is the process of identifying latent topics in a large set of text documents. It is an example of a NLP method, which examines unstructured text data collections to discover topics that are difficult or impossible to uncover by human efforts alone. The method has often been perceived as a relatively new development in information retrieval sciences.

Topic modeling methods such as Probabilistic Latent Semantic Analysis and Latent Dirichlet Allocation (LDA) have become increasingly popular (Blei, 2012; Hannigan et al., 2019). However, attempts to extract topics from unstructured text using Factor Analysis (FA) techniques can be found as early as the 1960s (Hecking & Leydesdorff, 2018). In all topic modeling methods, it is necessary to build a Document-Term Matrix (DTM) that contains the number of term occurrences per document. The rows of the DTM usually represent the documents and the columns represent the frequencies of the whole vocabulary of distinctive terms used in the documents. The use of the DTM approach in text analytics turn thematic analysis of texts into a quantitative problem based on analysing the relations between the frequencies of words used in each of the documents. The LDA approach to topic modeling uses probability theory to process the DTM (Blei, 2012), while the FA approach is deterministic and uses linear algebra methods (Hecking & Leydesdorff, 2018).

The applications of topic modeling in management research have recently been summarized by Hannigan et al. (2019). The topic modeling algorithm considers every webpage text as a text document that is a mixture of topics, and every topic as a word or mixture of words. Words can be shared between topics and the topics can be shared among abstracts. The algorithm identifies combinations of words that are semantically interrelated and tend to appear together across the different documents. The combinations of words help identify specific themes and patterns that are latently present in the corpus. In addition, the topic model organizes the corpus by clustering the documents corresponding to each topic. The documents associated with a given topic are ranked in terms of the degree of their association with it. A closer examination of the topical organization of the documents enables interpretation of the overall theme and labelling of the topics (Blei, 2012; Hecking & Leydesdorff, 2018; Hannigan et al., 2019).

The topic modeling approach has also been applied to examine the market offers of a sample of companies in a specific business technology sector (Mamosian et al., 2018). Mamosian et al. (2018) adopted the format and logic of the modeling approach to examine a corpus of text documents created by scraping the websites of approximately 140 Canadian photonics technology companies. The main value of the suggested approach comes from a simple practical insight —the fact that the names of the text documents associated with a specific topic could be used to identify the companies from the websites of which provided the texts of these same documents. We can therefore use the topic model to identify the groups of companies that are most highly associated with a specific topic. In this sense, Mamosian et al. (2018) show that topic modeling can be used as a way of categorizing samples of companies based on the information they provide on their websites. And, since most of the information provided by companies on their websites is about what they offer to their potential customers, the topic model will most probably categorize the companies in terms of their market offers. This is a valuable practical insight that was used as a methodological basis for the present study.

3. Methodology

Our research approach is based on the application of WordStat —a commercial tool using FA to perform topic modeling (Hecking & Leydesdorff, 2018). The topic modeling algorithm identifies the key terms and the sets of documents corresponding to each of the topics. The research method included several steps following the logic suggested by Mamosian et al. (2018). First, we identified the sample of digital health companies by choosing a list of the top 100 digital health companies in the world created by the Medical Futurist Organization (see Appendix A). The companies were selected based on four criteria: mindset for innovation, truly disruptive technology, viable business model, and clear dedication to digital health. The second step was scraping the information from their websites and cleaning the data to

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

be used in the topic modeling process. We used a commercial tool (https://firescraper.com/) to scrape all accessible company webpages, which resulted in a corpus of nearly 5,000 documents corresponding to the webpages of 92 companies (8 company websites were not able to be scraped, see Appendix A).

The corpus of documents was then subjected to our topic modeling approach to identify the emerging topics and the text documents associated with them. We used the Wordstat software, which enables research based on NLP, statistical, and feature analysis, together with multiple post-processing and visualization options (https://provalisresearch.com/products/content-analysis-software/). Wordstat performs topic modeling based on FA (Hecking & Leydesdorff, 2018). The topic extraction is achieved by computing a (word x document) frequency matrix, or alternatively by segmenting documents into smaller chunks and computing a (word x segment) frequency matrix. Once this matrix is obtained, a FA with Varimax rotation is computed to extract a small number of factors. All words with a factor loading value higher than 0.3 are then retrieved as part of the extracted topic.

Each of the topics is characterized with a set of words that tend to appear together and define an online communication theme that is common to many of the documents and thus, respectively, the companies. The documents were named in a way that they could be easily linked to the corresponding company names. Thus, we identified the companies associated with specific market offer types and examined the information on their websites to confirm our initial associations.

Table 1. List of topcis identified in the topic modeling process together with some of the most frequent words associated with them. The Coherence and Eigenvalue indicate the relevance of the specific topic within the entire topic model.

No	Topic	Most frequent topic words	Coherence	Eigenvalue	% Cases
1	Genetics	Variants; sequencing; genome; genetics; genes; factors; report; customers; testing; DNA test	0.72	8.92	12.76%
2	Diabetes	glucose; invasive; diabetics; continuous; flexible; affordable; developing; diabetes; company; monitor; diabetes prevention	0.71	8.05	8.96%
3	Side effects	Side effects; side effects and cost; treatment reviews; effectiveness major; effectiveness moderate	0.51	2.80	2.54%
4	Remote health care	Smart patients; online community; patient care services; smart patients website; share information; telemedicine; health systems	0.49	3.72	18.38%
5	Women & pregnancy	Natural cycles; birth control; hormonal birth control; birth control methods; pregnancy; ovulation; women; menstrual cycle; early pregnancy; fertile days; fertile window; pregnancy test; pregnant women; after ovulation; negative pregnancy test	0.43	3.18	4.24%
6	Privacy policy & personal information	Privacy; policy; personal; information; security; personal data; committed to protecting; terms; agreement; services	0.42	4.55	8.90%
7	University & medical centers	University; school; medicine; center; director; medical school; school of medicine; medical center; scientific; society; American society	0.41	5.73	6.74%

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

Table 1. List of topcis identified in the topic modeling process together with some of the most frequent words associated with them. The Coherence and Eigenvalue indicate the relevance of the specific topic within the entire topic model (cont'd).

8	Clinical trials	Clinical trials; clinical research; clinical studies; FDA cleared; clinically validated personal ECG solution; FDA clearance	0.40	2.58	12.56%
9	Virtual reality	Virtual reality; pain; therapy; virtual care; pain management; VR therapy; pain relief; anxiety	0.40	2.48	5.92%
10	Cancer	Skin cancer; cell carcinoma; ovarian cancer; cancer cells; lung cancer; breast cancer; basal cell carcinoma	0.40	2.98	5.44%
11	Vision	Vision; eye; vision test; vision test from home; personal vision tracker; eyeglass numbers	0.38	1.72	4.24%
12	Nutrition	Gluten; celiac; disease; diet; food; celiac disease; free diet; gluten free; food allergy; untreated celiac disease	0.38	3.07	3.68%
13	Hearing	Hearing loss; hearing screening; hearing health; hearing test; healthy hearing; hearing aids; hearing healthcare; access to hearing	0.38	2.79	2.00%
14	Mental health	Stress; mind; mental; practice; anxiety; mindfulness meditation	0.37	1.58	2.26%
15	Blood pressure	High blood pressure; monitor; blood test; blood oxygen; blood flow; oxygen levels; upper arm monitor; wrist monitor	0.34	1.97	2.12%
16	Asthma	Asthma; medication; asthma symptoms; people with asthma; asthma control; manage; respiratory disease; asthma patients	0.33	1.76	2.34%
17	Wearable devices	Wearable diagnostic devices; wearable technology; wearable sensors; subscription service; personalized lifestyle coaching programs; invasive glucose monitors	0.33	1.93	2.46%
18	Fitness	Activity; rate; sleep; tracking; heart rate; deep sleep; physical activity; quality sleep; heart rate monitor; heart rate variability	0.32	2.34	2.28%
19	Mobile apps	App; apps; users; google play; app store; mobile app; health apps; download the app; free app	0.32	1.62	1.92%
20	Stethoscope	Stethoscope; cardiac; digital stethoscope; electronic stethoscope; core digital stethoscope; heart murmurs; heart sounds	0.31	1.66	1.10%
21	Ultrasound	Ultrasound; lung ultrasound; handheld ultrasound; ultrasound system; portable ultrasound; point of care; ultrasound scanners; ultrasound education; scanner	0.30	1.85	1.86%
22	Artificial intelligence	Artificial intelligence; therapeutic artificial intelligence; human intelligence; machine learning; reinforcement learning	0.28	1.74	4.88%
23	Response to Covid	Covid; pandemic; response to Covid	0.25	1.88	3.06%

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

The final analysis included a discussion of the ability of the method to distinguish between topics related to specific market offers or application domains (including the specific companies focusing on them) vs. topics related to issues discussed online by the majority of the companies. This analysis allowed us to develop insights that could help automate the analytical approach and enhance its ability to work with larger samples of companies.

4. Results

The application of the topic modeling approach on the corpus of approximately 5,000 text documents associated with the 92 innovative digital health companies resulted in categorizing the dominant market offer types in the sample and the issues that companies found necessary to communicate online. It is important to note that "dominant market offer type" can also refer to the specific application domain, which could incorporate different but similar market offers.

The post-processing of the topic modeling results identified 23 topics (Table 1) and the companies that are most highly associated with each of the topics. Table 1 includes the topic labels (second column) and the set of most frequent words corresponding to a topic (third column). The Coherence value of the topics is presented in the fourth column. It measures the degree of semantic consistency of the most frequent words in the topic. The Coherence value helps in distinguishing between topics that are thematically interpretable in the context of the specific study, and topics that are artifacts of statistical inference but appear to be irrelevant to the context. The fifth column of Table 1 shows the Eigenvalue of each topic. These are the eigenvalues of the factors derived by means of the FA method and should be larger than 1. The last column (% Cases) shows the percentage of the documents that are most highly associated with a given topic.

The Coherence and Eigenvalue indicate the relevance of the specific topic within the entire topic model.

The topics presented in Table 1 are ranked according to their Coherence value. The selection of exactly 23 topics was based on the choice of a threshold Coherence value of 0.25. This threshold value ensures a reasonable semantic consistency of the topic words and topic eigenvalues larger than 1.

The scope of this article does not allow us to describe the full details of the fine structure of the topic model presented in Table 1. The rest of this section will concentrate on describing selected topic results to demonstrate the ability of the method and identify 2 main types of topics, focusing on: i) a specific digital health market offer type or application domain (for example, genetics, diabetes, blood pressure, vision, artificial intelligence); b) an issue or an online communication theme shared by the majority of the firms (for example, remote health care, response to Covid-19). The set of results shown in Figures 1-7 below were chosen as a concise way of providing some

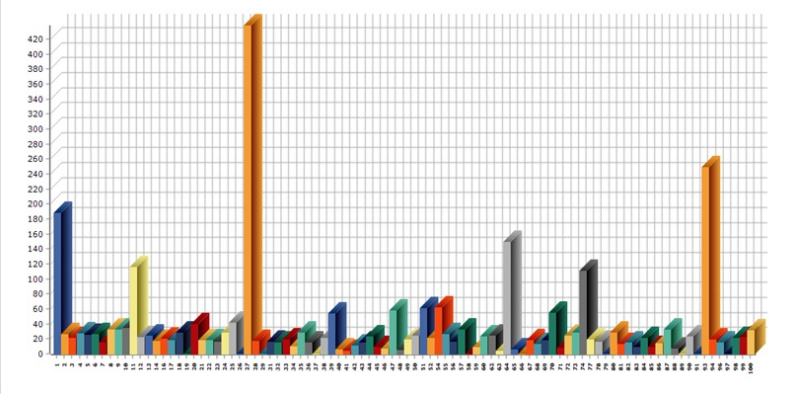

Figure 1. Topic 1 (Genetics) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe). The numbers on the horizontal axe of the graph correspond to the numbering of the list of companies provided in Appendix A. The graph helps in identifying the businesses that are most highly associated with Genetics.

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

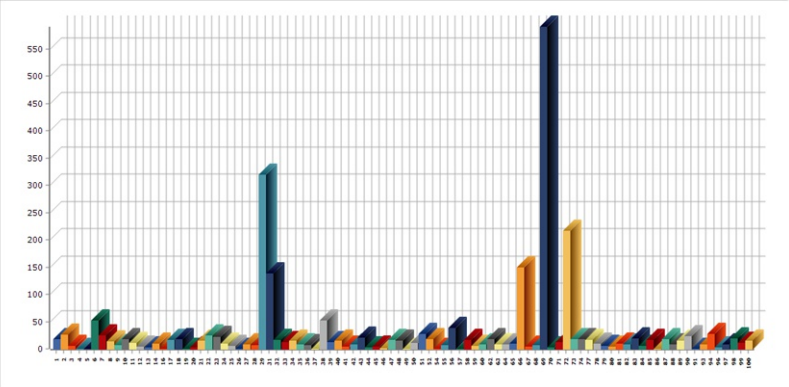

Figure 2. Topic 2 (Diabetes) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe). The graph helps in identifying the businesses that are associated with diabetes issues.

representative examples of these three types of topics.

Fig. 1 shows the Topic 1 (Genetics) *Rate per 10,000 words* (Rp10KW) for each of the companies in the sample. RP10KW is a variable corresponding to each company's frequency of use of the words associated with the topic Genetics (see the first row of Table 1).

Fig. 1 shows that 5 companies that are most highly associated with the Genetics topic. They correspond to the columns that are relatively higher than the rest of the data. To identify these companies, we have looked for companies with Rp10KW value larger than 120. The choice of the threshold value 120 for the Rr10KW variable was made in a way to allow identifying a manageable number of companies that are most highly associated with the specific topic. Identifying these companies meant that they use genetics-related words on their websites more frequently than the rest of the companies.

The 5 companies are as follows (provided in the priority of their degree of use of the Genetics topic words; the numbering follows the list provided in Appendix A):

- # 27, *Dante Labs*: Takes a proactive approach to health with insights on genetic predispositions, drug response and well-being.
- # 93, *Veritas*: Operates a high complexity next generation sequencing (NGS) laboratory.
- # 1, *23andme*: A genetic and health company involving scientific, data, and genetic insights.
- # 64, *MyDNA*: Leverages the world's leading genotyping technology and scientific algorithms to learn more about DNA.
- # 11, *Atlas Biomed*: Provides a two-feature DNA test (with a detailed genetic profile of one's health, nutrition, physical activity and geographical ancestry) and Microbiome test (analyzing the types of bacteria present and their proportion in the overall microbiome).

The brief description of the focus of the 5 companies shows that they are indeed dealing with a business related to genetics. Fig. 1 is quite representative because it shows that the topic modeling approach can spot companies that are predominantly dedicated to a specific application domain (in this case genetics): next generation sequencing, DNA testing, developing genetic predisposition insights, etc. Our visual examination of Fig. 1 suggests that it could be used as a basis for developing a fully automated web search and business analytics process that would enhance the abilities of both scholars and practitioners to examine the dominant offers and communication priorities of companies in specific business sectors.

Fig. 2 shows that 5 companies are most highly associated with the Diabetes topic. To identify these companies, we have used again an Rp10KW threshold value of 120. The 5 companies are as follows:

- # 69, *Nemura Medical*: Produces a variety of devices

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

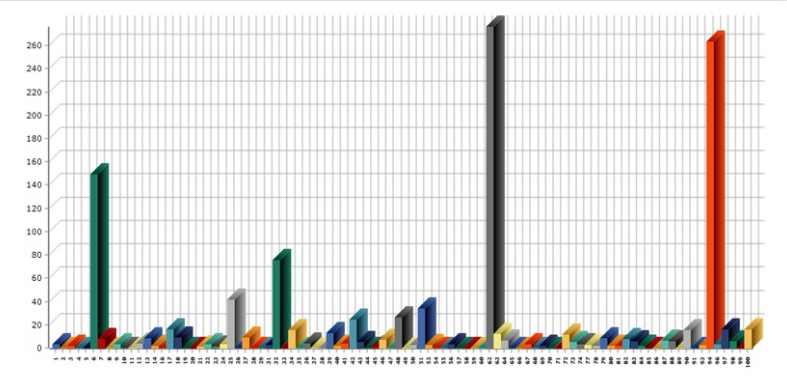

Figure 3. Topic 15 (Blood pressure) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe). The graph helps in identifying the businesses that are associated with blood pressure issues.

to monitor glucose and weight loss, athletes & viral infection.

- # 29, *Dexcom*: Produces devices for monitoring glucose level.
- # 72, *Omada Health*: Produces a device and application for continuous glucose monitoring.
- # 31, *Diabeloop*: Produces devices for real time monitoring of type 1 diabetes.
- # 66-*MySugr:* Markets app to monitor diabetes, diabetes management kits, and coaching related to diabetes.

The brief description of the focus of the 5 companies above shows that they are indeed dealing with a business related to diabetes.

Fig. 3 shows that 3 companies are most highly associated with the Blood pressure topic (Rp10KW threshold value > 120). The 5 companies are as follows:

- # 62, *Mocacare*: Provides devices to monitor blood pressure.
- # 94, *Viatom*: Designs and manufactures healthcare products (wearable and home medical devices) such as pulse oximeters, portable vital signs monitors, EKG/ECG Holter Monitor and portable blood pressure monitor.
- # 6, *Alivecor*: Delivers intelligent, highly personalized

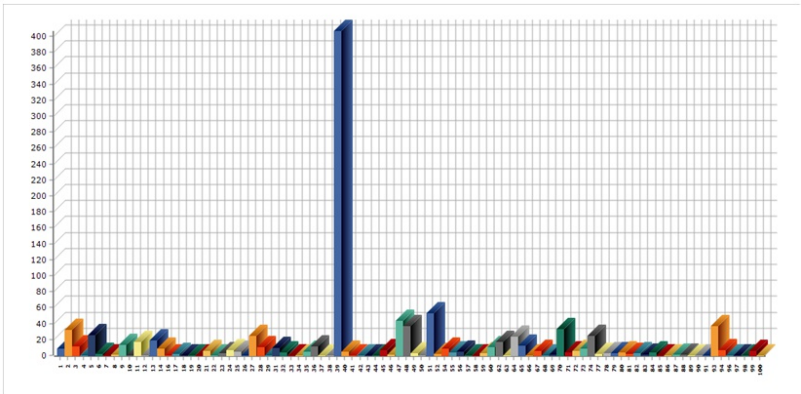

Figure 4. Topic 11 (Vision) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe). The graph helps in identifying the one company that is most highly associated with vision issues.

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

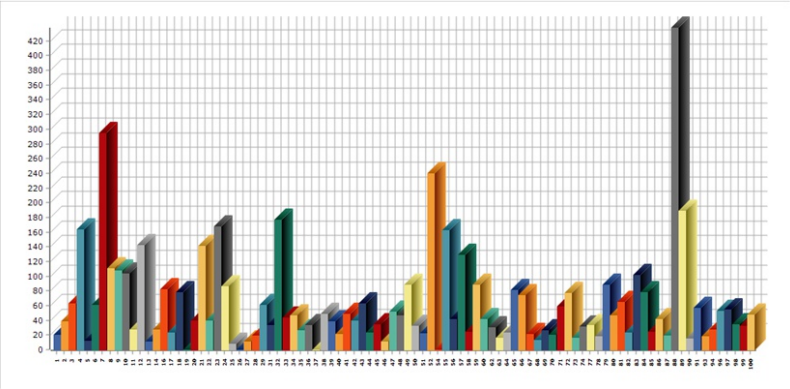

Figure 5. Topic 4 (Remote health care) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe). The graph helps in identifying the companies that are referring most frequently to Remote health care. However, the graph is qualitatively different from the ones shown above since it shows that the topic is discussed by a relatively larger number of companies compared to the previous cases shown in Figures 1-4.

heart data with advanced determinations.

The brief description of the focus of these 3 companies above shows that they are indeed dealing with a business related to Blood pressure issues.

Fig. 4 shows that there is one company that is most highly associated with vision - # 39, Eyeque, which provides eye care solutions with self-administered eye test devices.

Fig. 5 shows that there are 10 companies discussing more intensively remote health care issues on their websites (Rp10KW threshold value > 120). The top 5 companies associated with this topic are as follows:

- # 88, *Smart Patients*: An online community for patients and families affected by a variety of illnesses.
- # 7, *AmWell:* Market services to accelerate the patient journey combined with opportunities to scale telehealth.
- # 52, *INTouch Health*: Provides a virtual care platform and telehealth devices.
- # 4, *AdhereTech*: Provides a software platform analyzing a number of data feeds, such as: real-time adherence information from our devices, patient messages & feedback, and pharmacy inputs & data.
- # 32, *EKO*: Provides a platform that brings together advanced stethoscopes, patient and provider software, and AI-powered analysis with pharmacy inputs & data.

The visual examination of Fig. 5 shows that the Remote health care topic is both qualitatively and quantitatively different from the previous examples because it is shared among a relatively larger number of companies. Fig. 5 is an example of a "noisier" visual pattern that could be clearly differentiated from the previous ones (Figures 1-4) because of the higher-level background of companies that are less associated with the specific topic. Such a visual pattern can be associated with topics that are not related to a specific application domain, but rather to a particular common feature of many different products developed by a larger number of companies.

Figure 6 shows that there are many companies that are frequently referring to the Covid pandemic with an average intensity that is lower compared to the previous topics. The Rp10KW value of all these companies is visibly lower than the threshold value of 120 that was used in the previous cases. The visual pattern of the graph suggests that the Covid topic may not be associated with a specific application domain, but instead with an issue that is discussed by most of the companies in the sample.

The top 2 companies associated with the Artificial intelligence topic are as follows:

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

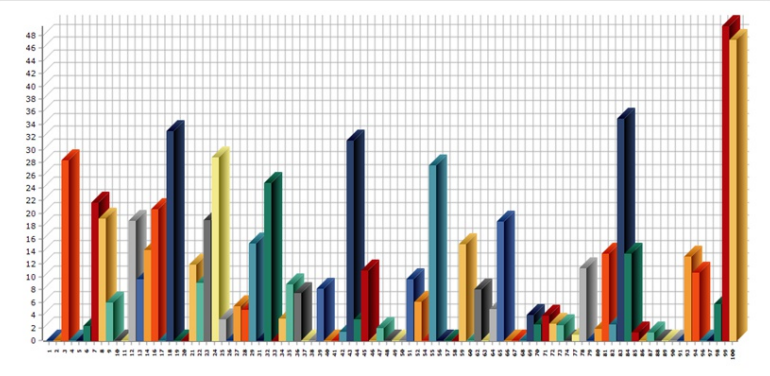

Figure 6. Topic 23 (Response to Covid) *Rate per 10,000 words* (vertical axe) for each of the companies in the sample (horizontal axe).

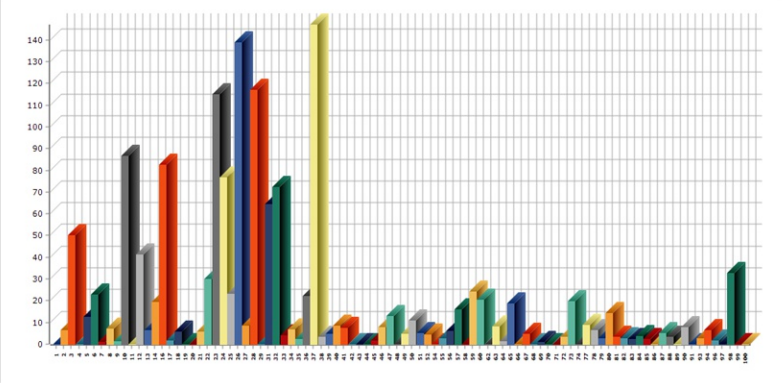

Figure 7. Topic 22 (Artificial intelligence) *Rate per 10,000 words* (vertical axis) for each of the companies in the sample (horizontal axis). The graph helps in concluding that there are more than 5 companies developing artificial intelligence-based products.

- # 26, *Cyberdyne*: Develops a variety of innovative cybernics devices and interfaces and advanced AI-Robot products to enable early detection and prevention for health maintenance, to improve our aging workforce, and to respond to the shrinking workforce.

- # 28, *DeepMind*: Develops safe artificial intelligence systems to provide intelligence solutions and advance scientific discovery for all.

The identification of the two companies was based on the Rp10KW threshold value of 120. However, Fig. 7 shows that the Artificial intelligence topic is associated with a larger number of companies using AI as an enabling technology in different types of market offers or business applications (the value of their Rp10KW is very close to the threshold value). At the same time, the visual pattern of the graph shown in Fig. 7 is different from the "noisy" patterns characterizing the companies associated with the Remote health care and Response to Covid topics. It allows for two potential ways of interpreting the AI topic - as an application domain or as a specific technology that is used in companies' products and services.

The analysis provided here could be extended to the rest of the 23 topics shown in Table 1. A closer examination of the description of the topics will show they that could be categorized in terms of the two main types we identified above - a specific digital health application domain or an online communication theme. The results suggest that we can identify fifteen application domains: Genetics, Diabetes, Women & pregnancy, Virtual reality, Cancer, Vision, Nutrition, Hearing, Mental health, Blood pressure, Asthma, Fitness, Stethoscope, Ultrasound, and Artificial intelligence. We can also identify eight online communication themes that are common to most of the

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

companies: Side effects, Remote health care, Privacy policy & personal information, Clinical trials, University & medical centers, Wearable devices, Mobile apps, and Response to Covid. Some of these themes refer to medical issues such as side effects and clinical trials, while others refer to product-related issues or technologies such as wearability and mobility.

5. Conclusion

The above article summarized the results of a research study adopting a text analytics (topic modeling) approach to identifying dominant themes discussed on the websites of a sample of 100 innovative digital health companies. Two major types of topics were identified - market offer types or business application domains, and issues of concern to many companies in the sample that are not directly associated with their specific market offers. Our approach allowed identifying companies that are most highly associated with specific application domains. The combination of business application domains and their most representative companies could be used to map the business focus of innovative companies in the digital health sector. The specific results enhance our understanding of what innovative digital health companies offer, along with their online thematic focus. However, the value of the suggested analytical approach consists in its ability to be replicated in the context of other business sectors. Future studies may focus on developing more powerful post-processing text analytics capabilities that could deepen the insights provided by the topic model.

The article does not pretend to make a specific theoretical contribution. Its main contribution is methodological since it demonstrates the potential of a new application of the topic modeling approach to generate innovation management research insights. The instrumentalization of the methodological approach adopted in this study represents an opportunity to develop valuable insights not only for busines innovation scholars and executive managers of new ventures, but also for managers of incubators, accelerators, and innovation centers interested in enhancing their business intelligence services to their clients. In addition, it could benefit investors looking for promising business areas, or policy makers looking to promote certain types of businesses, as well as ethicists looking to understand similarities and differences in new technology businesses associated with emerging ethical dilemmas, for example, AI and genetics.

References

Blei, D. 2012. Probabilistic Topic Models. Communications of the ACM, 55(4): 77-84. DOI: https://doi.org/10.1145/2133806.2133826

Hannigan, T., Haans, R., Vakili, K., Tchalian, H., Glaser, V., Wang, W., Kaplan, S. & Jennings, P. 2019. Topic Modeling in Management Research: Rendering New Theory from Textual Data. *Academy of Management Annals*, 13(2): 586-632. DOI:

Hecking, T. & Leydesdorff, L. 2018. Topic Modelling of Empirical Text Corpora: Validity, Reliability, and Reproducibility in Comparison to Semantic Maps: https://arxiv.org/abs/1806.01045.

Johnson, M., Christensen, C. & Kagermann, H. 2008. Reinventing your Business Model. *Harvard Business Review*, December, 51-59. DOI:

Mamosian, H., Tanev, S., Bailetti, T. & Tzolov, V. 2018 Using Online Text Analytics to Differentiate the Market Offers of Technology Firms. *Proceedings of the ISPIM Connects Fukuoka Conference*, Iain Bitran et al., Eds., LUT Scientific and Expertise Publications, Dec. 2-5, 2018, Fukuoka, Japan.

Tanev, S., Liotta, G. & Kleismantas, A. 2015. A Business Intelligence Approach Using Web Search Tools and Online Data Reduction Techniques to Examine the Value of Product-Enabled Services. *International Journal Experts Systems with Applications,* 42(21): 7582-7600.

Tang, J., Kacmar, K.M., & Busenitz, L. 2012. Entrepreneurial Alertness in the Pursuit of New Opportunities. *Journal of Business Venturing* 27: 77-94.

Wulfovich, S. & Meyers, A., Eds. 2020. *Digital Health Entrepreneurship*, Springer.

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

About the Authors

Renée Emby, B.A, MABA, is a Technical Advisor with Shared Services Canada, Ottawa, ON, Canada. Renée currently leads a team of employees as they deliver Information Management and Information Technology services to Canadians and the Government of Canada. Before working with Shared Services Canada, Renée was employed with Canada Border Services Agency where she worked in business analytics pertaining to national and international security. Renée began her academic journey at Carleton University where she obtained an undergraduate degree in Business Law (2020) and a Masters of Applied Business Analytics (2021). Renée is continuing her education at the University of Ottawa, where she is pursing a Certificate in Business Process Improvement (2021). Renée's experience and interests pertain to national security, security of information, data analytics, service management and delivery.

Daman Arora, is a Software Engineer, currently working towards a Master of Applied Business Analytics degree in TIM Program at Carleton University, Ottawa, ON, Canada. Prior to that, Daman studied Computer Systems Technician program (2017, Algonquin College, Ottawa, ON, Canada) and worked as an Intern as well as a Full Time Software Engineer in the Cloud and Cognitive Support business unit of IBM Canada. Daman has a keen interest in the field of Cloud Computing, DevOps, Data Analytics, & Machine Learning. Daman also enjoys contributing to Open Source projects and has made significant code, and non-code contributions to various projects, notably, Kubernetes, TrinoDB, & Apache CloudStack. Daman Arora is member of the Inaugural Class of Community Advocates at Ambassador Labs for the period of 2021-2022. Daman is continuing his education at York University, where he is pursuing a Certificate in DevOps (2021).

Madiha Rehman is a master's degree holder of Business Analytics associated with Technology Innovation Management from Carleton University, Ottawa, ON, Canada (2021). Before that Madiha did her Honors in Bachelor of Computer Science (2002). Madiha is currently working as a Technical Support Engineer and a Business Development Representative. Madiha is a tech-savvy professional skilled in many areas as an analyst, technical support provider, customer support and success and business development. Madiha is a committed professional team player who has passion for continuous improvement and learning and willingness to take on new and advanced roles. Madiha would like to pursue a highly rewarding career, where she can utilize her skills and knowledge, develop new skills and contribute in the accomplishment of organizational goals.

George Tanev, MSc, MEng, is a Product Owner at Export Development Canada in Ottawa, ON, Canada. He works in innovating and developing knowledge based solutions to support Canadian companies go and grow global. George's background spans multiple interdisciplinary fields including systems engineering, medical device research and development, and entrepreneurship. George's academic backround includes a BEng in Biomedical and Electrical Engineering (Carleton University, 2008), a MEng in Medicine and Technology (Technical University of Denmark, 2012), and a MSc in Technology Innovation Management (Carleton University, 2021). George's research interests include applied business analytics, medical technologies, product innovation and cybersecurity.

Abdulla Aweisi, MEnt, B.Sc., currently is working as IT Manager with TechBrew Robotics, Salmon Arm, BC, Canada. Abdulla has more than 15 years of experience in the Information Technology field, with a demonstrated history of working in the Building Materials Manufacturing \ Retail industry. Skilled in IT Digital & Business Transformation, Business Processes re-engineering, ERP Implementations, and IT Strategy. Passionate about Business Intelligence, Data Science, and Entrepreneurial Ecosystems. Holding a B.Sc. in Computer Science (2006) from Princess Sumaya University for Technology (PSUT), Amman, Jordan, and Masters of Entrepreneurship, Technology Innovation Management (TIM) (2021) from Sprott School of Business, Carleton University, Ottawa, ON, Canada.

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

Appendix A. List of the digital health companies included in the sample (companies in red did not allow scraping their online text content).

No	Company name	Web
1	23andme	https://www.23andme.com/en-ca/
2	Acurable	https://acurable.com/
3	Ada Health	https://ada.com/
4	AdhereTech	https://www.adheretech.com/
5	AioCare	https://aiocare.com/patient/
6	Alivecor	https://www.alivecor.com/
7	AmWell	https://business.amwell.com/
8	Antidote	https://www.antidote.me/
9	AppliedVR	https://appliedvr.io/
10	Arterys	https://arterys.com/
11	Atlas Biomed	https://atlasbiomed.com/ca
12	Augmedix	https://augmedix.com/
13	Ava Health	https://www.avawomen.com/
14	Babylon Health	https://www.babylonhealth.com/
15	BeatO	https://www.beatoapp.com/
16	Behold.AI	https://behold.ai/
17	Bonzun	https://bonzun.com/
18	Butterfly Network	https://www.butterflynetwork.com/
19	Calm	https://www.calm.com/
20	Cellink	https://www.cellink.com/
21	Circulation	https://www.circulation.com/
22	Clarius	https://clarius.com/
23	CloudMedX	https://cloudmedxhealth.com/
24	Corti.AI	https://www.corti.ai/
25	Curiscope	https://www.curiscope.com/
26	Cyberdyne	https://www.cyberdyne.jp/english/
27	Dante Labs	https://www.dantelabs.com/
28	DeepMind	https://deepmind.com/
29	Dexcom	https://www.dexcom.com/en-CA
30	D-Eye	https://www.digitaleyecenter.com/
31	Diabeloop	https://www.diabeloop.com/
32	EKO	https://www.ekohealth.com/
33	Ekso	https://eksobionics.com/
34	Ekuore	https://www.ekuore.com/
35	Elvie	https://www.elvie.com/
36	Empatica	https://www.empatica.com/en-int/
37	Enlitic	https://www.enlitic.com/
38	etectRX	https://etectrx.com/
39	Eyeque	https://www.eyeque.com/
40	FabRX	https://www.fabrx.co.uk/
41	Firsthand Technology	https://firsthand.com/
42	Fitbit	https://www.fitbit.com/global/en-ca/home
43	Florence	https://florencehc.com/
44	Garmin	https://www.garmin.com/en-CA/
45	Ginger.IO	https://www.ginger.com/
46	Headspace	https://www.headspace.com/
47	Hearscope	https://www.hearxgroup.com/hearscope/
48	HeraBeat	https://herabeat.com/
49	Hocoma	https://www.hocoma.com/us/
50	iDoc24	https://www.idoc24.com/
51	Imaware	https://www.imaware.health/
52	INTouch Health	https://intouchhealth.com/
53	Intuitive Surgical	https://www.intuitive.com/en-us
54	Liftware	https://www.liftware.com/
55	Maven	https://www.mavenproject.org/
56	MC10	https://mc10inc.com/
57	Medical Realities	https://www.medicalrealities.com/
58	Medtronic	https://www.medtronic.com/ca-en/index.html
59	Medwand	https://ces.tech/
60	Microsoft Hololens	https://www.microsoft.com/en-us/hololens/industry-healthcare
61	Misfit	http://www.misfit.com/
62	Mocacare	https://mocacare.com/
63	Muse	https://choosemuse.com/
64	MyDNA	https://www.mydna.life/
65	MySense.AI	https://www.mysense.ai/
66	MySugr	https://www.mysugr.com/en/
67	Natural Cycles	https://www.naturalcycles.com/
68	Natural Machines	https://www.naturalmachines.com/
69	Nemura Medical	https://nemauramedical.com/
70	Nima	https://nimasensor.com/
71	Not Impossible Labs	https://www.notimpossible.com/
72	Omada Health	https://www.omadahealth.com/
73	Omron	https://www.omron.com/global/en/
74	Oncompass Medicine	https://www.oncompassmedicine.com/
75	OneRemission	http://oneremission.com/#/
76	Oscar Health	https://www.hioscar.com/
77	Osso VR	https://ossovr.com/

Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies

Abdulla Aweisi, Daman Arora, Renée Emby, Madiha Rehman, George Tanev and Stoyan Tanev

Appendix A. List of the digital health companies included in the sample (companies in red did not allow scraping their online text content). (cont'd)

78	Oura	https://ouraring.com/
79	Patients Like Me	https://www.patientslikeme.com/
80	Philips	https://www.philips.ca/c-e/hs/hearing-and-hearing-loss.html
81	Pilleve	https://www.pilleve.com/
82	Polar	https://www.polar.com/ca-en
83	Propeller Health	https://www.propellerhealth.com/
84	Psious	https://psious.com/
85	Rewalk Robotics	https://rewalk.com/
86	Sensely	https://www.sensely.com/
87	Skinvision	https://www.skinvision.com/
88	Smart Patients	https://www.smartpatients.com/
89	Surgical Theater	https://surgicaltheater.net/
90	Thryve	https://www.thryveinside.com/
91	Urbandroid	http://team.urbandroid.org/projects/
92	Veebot	http://www.veebot.com/
93	Veritas	https://www.veritasgenetics.com/
94	Viatom	https://www.viatomtech.com/
95	Virtually Better	https://www.virtuallybetter.com/
96	Wahoo	https://www.wahoofitness.com/
97	Withings	https://www.withings.com/ca/en/
98	Woebot	https://woebothealth.com/
99	Xenex	https://xenex.com/
100	Zipline	https://flyzipline.com/

Author Bios (cont'd)

Stoyan Tanev, PhD, MSc, MEng, MA, is Associate Professor of Technology Entrepreneurship and Innovation Management associated with the Technology Innovation Management (TIM) Program, Sprott School of Business, Carleton University, Ottawa, ON, Canada. Before re-joining Carleton University, Dr. Tanev was part of the Innovation and Design Engineering Section, Faculty of Engineering, University of Southern Denmark (SDU), Odense, Denmark. Dr. Tanev has a multidisciplinary background including MSc in Physics (Sofia University, Bulgaria), PhD in Physics (1995, University Pierre and Marie Curie, Paris, France, co-awarded by Sofia University, Bulgaria), MEng in Technology Management (2005, Carleton University, Ottawa, Canada), MA in Orthodox Theology (2009, University of Sherbrooke, Montreal Campus, QC, Canada) and PhD in Theology (2012, Sofia University, Bulgaria). Stoyan has published multiple articles in several research domains. His current research interests are in the fields of technology entrepreneurship and innovation management, design principles and growth modes of global technology start-ups, business analytics and text mining. He is also interested in interdisciplinary issues on the interface of science and theology.

Citation: Aweisi, A., Arora, D., Emby, R., Rehman, M., Tanev, G., and Tanev, S. 2021. Using Web Text Analytics to Categorize the Business Focus of Innovative Digital Health Companies. Technology Innovation Management Review, 11 (7/8): 65-78.

http://doi.org/10.22215/timreview/1457

Keywords: Digital health sector, topic modeling algorithm, market offer, value proposition, machine learning, web analytics