NLP for Market and Competitive Intelligence

Igor Menghini[1]

[1]Roche Diagnostics S.p.A., Viale G.B. Stucchi 110, 20900 Monza, Italy

Abstract

Market intelligence plays a crucial role in informing strategic decision making in the competitive biomedical sector. With the advancement of digital technologies, there is growing interest in leveraging these tools to enhance traditional market intelligence approaches. This article presents a case study that explores the practical applications of natural language processing (NLP), and advanced analytic methods in the context of market and competitive intelligence. The case study focuses on the use of these digital tools to monitor conferences and analyze competitor hiring trends. The findings highlight the potential of digital technologies to provide valuable insights for evidence-based decision making in the business domain.

Keywords

NLP, Text analysis, Market intelligence, Competitive intelligence, Social media

1. Introduction

The biomedical industry is characterized by rapid technological advancements, intense competition, and evolving market dynamics. In this fast-paced environment, market intelligence plays a pivotal role in helping organizations understand trends, monitor competitors, identify new opportunities, and make informed strategic decisions.

Traditionally, market intelligence has relied on methods such as primary and secondary research, surveys, and expert interviews [1], however, the digital era has unlocked vast amounts of publicly available online data that can potentially augment traditional approaches when leveraged through advanced digital technologies. This article presents a case study that explores the practical applications of digital tools, NLP, and advanced analytic methods, in the context of market and competitive intelligence.

2. Monitoring conferences via social media:

Scientific conferences, along with publications, serve as a key source of knowledge about trends in a particular domain, however, due to budget constraints or travel restrictions, physical attendance may not always be feasible. To overcome this limitation, we developed a scalable process to gather conference-related data from social media platforms, using relevant conference hashtags. (fig. 1)

3rd Italian Workshop on Artificial Intelligence and Applications for Business and Industries - AIABI, November 09, 2023, Rome, Italy

igor.menghini@roche.com (I. Menghini)

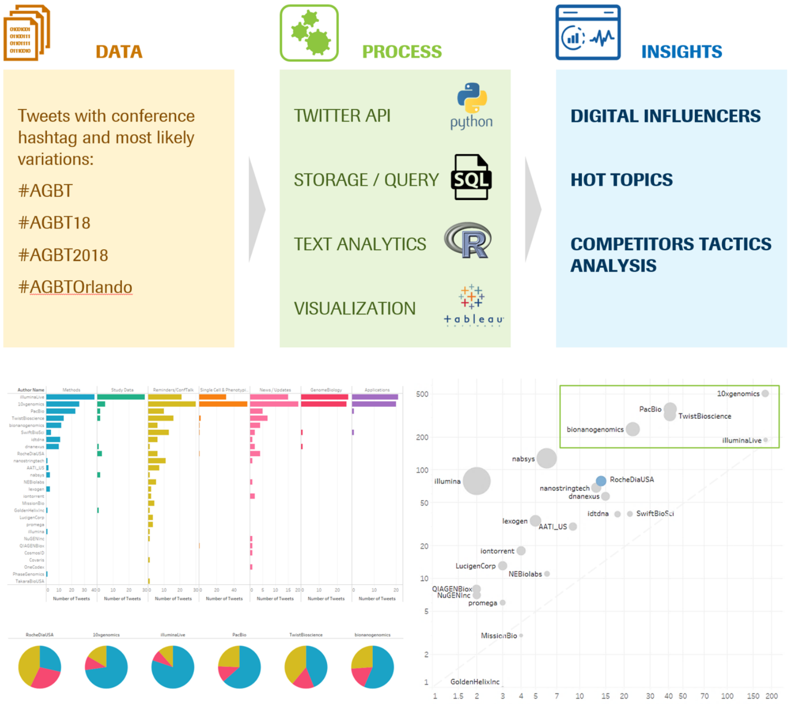


Figure 1: Conceptual pipeline and selected results from Twitter analysis

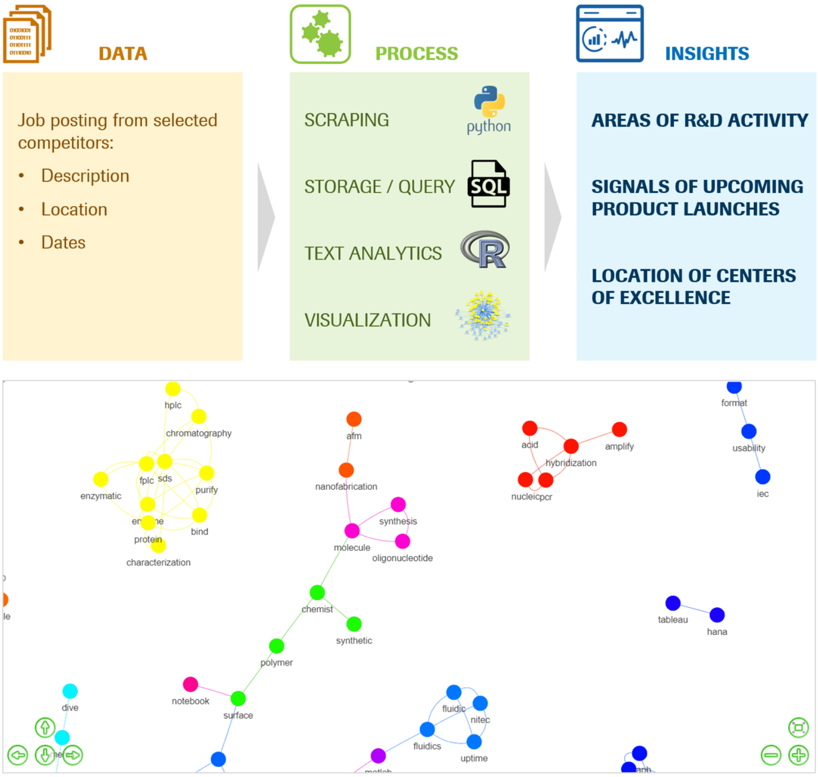
The analysis of social media data, in particular ”Twitter” (now ”X”) data, has become common practice during major global events [2, 3]. The real-time availability of large textual data-sets connected to specific events, often labeled through #hashtags, provides excellent corpora for topic extraction, exploration of social dynamics, and sentiment analysis. Following a similar principle, we explored the monitoring of medical and scientific conferences through social media to extract market and competitive intelligence.

We started our investigation by collecting historical data on a selected number of conferences to assess feasibility. During this initial phase, we experimented with different parameters and methods, gathering qualitative feedback from subject matter experts (SMEs) who had attended those events. Through our iterative and collaborative approach, we were able to identify several crucial factors that significantly impacted the success of our analysis. These factors encompassed the choice of hashtags, the minimum threshold for number of tweets, the parameters utilized for data processing, and the most informative visualizations.

Using NLP techniques and influence scoring algorithms, we identified influential discussants and discussion topics. Network graphs were then generated to visualize the relationships between key stakeholders, conferences, topics, and competitors. This approach provided a cost-effective means of virtually attending events and tracking competitor activities in real-time.

3. Competitive intelligence from job listings

Job postings analysis has been successfully used to assess labour market dynamics[4, 5, 6]. Focusing on postings from a specific competitor, or a limited number of key players, can provide valuable insights into an organization's strategic focus and can serve as an indicator of emerging trends in the biomedical industry.



DATA

Job posting from selected competitors:

- Description
- Location
- Dates

PROCESS

SCRAPING

STORAGE / QUERY

TEXT ANALYTICS

VISUALIZATION

INSIGHTS

AREAS OF R&D ACTIVITY

SIGNALS OF UPCOMING PRODUCT LAUNCHES

LOCATION OF CENTERS OF EXCELLENCE

Figure 2: Conceptual pipeline and selected results from job listing analysis

Through the use of web scraping and data extraction techniques, we collected a vast amount of job postings from various sources, including competitor's websites and job listing portals. The collected data was then processed using natural language processing (NLP) techniques to extract relevant information and identify patterns.

One of the primary challenges in analyzing job postings is the wide range of information contained in job descriptions: Beside specific roles, responsibilities, qualifications, and requirements, there is a variety of statements about the company culture, its mission and social responsibility, equality and inclusivity disclaimers, as well as generic corporate jargon with low information value.

To visualize the findings and facilitate interpretation, we developed interactive network graphs (fig. 2), which allowed our subject matter experts (SMEs) to identify connections and relationships between different job postings, skills, and competencies. By representing these relationships in a visual format, our SMEs were able to explore the information space and provide commentary and qualitative insights that enriched the contextual understanding of our competitors' strategy.

The insights derived from the analysis of competitor job postings provided us with valuable information for strategic decision-making. By understanding the skills and qualifications sought by competitors, we were able to identify potential gaps in our own talent pool and make informed decisions regarding recruitment and talent acquisition strategies. Additionally, the analysis shed light on emerging technology trends, validating our assumptions surrounding upcoming technology platforms pursued by our competitors.

4. Methods

The case study presented in this article utilized a multi-step process that combined various technologies. While the two pipelines used for conference intelligence and for job intelligence present some differences, they share a common logical structure:

4.1. Data collection

Custom scripts were developed and deployed on an cloud based virtual machine, daily triggers were set for automated data collection. Twitter data was retrieved via the twitter API, job posting data was obtained via web scraping. Data was saved and stored in a database with relevant metadata for efficient retrieval.

4.2. Data pre-processing

For tweets, datasets were defined as collection of entries related to the same conference, for job postings, datasets were defined as collection of jobs posted by the same company in a given time interval.

All the text elements were normalized to English using the Google Translate API and duplicate entries were removed, regular expressions were used to remove Twitter handles, URLs, and the conference #hashtags from tweets, and to remove a list of manually curated corporate statements from job descriptions. The pre-processed data was finally tagged as ready for analysis.

4.3. Data processing

NLP pipelines were developed and executed in R using the package tm[7].

After tokenization, a document-term-matrix (DTM) was created and normalized using the term-frequency inverse-document-frequency (TF-IDF) method, terms with frequency below 0.2% were then removed to reduce the matrix dimensions and its sparsity.

For tweets, we performed a spherical-kmeans clustering of the document-term-matrix using CLUTO[8] with three distinct partition numbers (k=6, k=7, k=8). We produced three separate output tables summarizing the most common terms in each partition, leaving to SMEs the choice of which parameter generated the most meaningful set of topics. We also calculated two user-centric scores for each conference, one as raw count of tweets per user, one as average number of re-tweets per tweet across the dataset.

For job postings, we calculated Spearman correlation between terms in the document-term-matrix and, leveraging the package igraph[9], constructed an undirected graph for each company. In these graphs, each node represented a term, whereas edge values were proportional to the correlation coefficients.

4.4. Data visualization

Results from each processing pipeline were stored in predetermined database tables and made accessible for visualization.

To display the conference analysis results, we chose to use a collection of standardized visualizations via Tableau. This decision was based on the tool's familiarity to our subject matter experts (SMEs) and its existing integration with our business processes.

For the job posting analysis, we utilized the visNetwork framework, which enabled SMEs to easily prune the graphs by selecting cut-off values through the user interface. This functionality allowed them to determine suitable parameters for the specific portion of the graph they were investigating.

5. Conclusions

In the rapidly evolving biomedical industry, market intelligence plays a crucial role in informing strategic decision-making. With the advent of digital technologies, organizations must evolve and adopt technology driven methods to enhance traditional market and competitive intelligence approaches.

This article highlights two use cases that demonstrate the effectiveness of NLP and other advanced methods in analyzing large amounts of data, extracting valuable insights, and guiding strategic decision-making.

A crucial component of our approach was the iterative improvement of our analysis pipelines in collaboration with subject matter experts. Their expertise and domain knowledge allowed us to rapidly evolve our methods, optimizing our processes not only for accuracy and reproducibility, but also for greater explainability of the results. Furthermore, the ability to alter a selected number of analysis parameters, simplified our method development, removing the need to precisely define the boundaries of applicability of our pipelines, while simultaneously instilling trust in the system.

References

[1] F. And, N. Sewdass, J. Calof, Contemporary Practices of Intelligence Support for Competitiveness, Foresight and STI Governance 14 (2020) 30–39. URL: http://creativecommons.org/licenses/by/4.0/. doi:10.17323/2500-2597.2020.3.30.39.

[2] C. Chew, G. Eysenbach, Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak, PLOS ONE 5 (2010) e14118. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0014118. doi:10.1371/JOURNAL.PONE.0014118.

[3] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hai, Z. Shah, Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study, J Med Internet Res 2020;22(4):e19016 https://www.jmir.org/2020/4/e19016 22 (2020) e19016. URL: https://www.jmir.org/2020/4/e19016. doi:10.2196/19016.

[4] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR) 34 (2002) 1–47. URL: https://dl.acm.org/doi/10.1145/505282.505283. doi:10.1145/505282.505283.

[5] P. G. Lovaglio, M. Cesarini, F. Mercorio, M. Mezzanzanica, Skills in demand for ICT and statistical occupations: Evidence from web-based job vacancies, Statistical Analysis and Data Mining: The ASA Data Science Journal 11 (2018) 78–91. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/sam.11372https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11372https://onlinelibrary.wiley.com/doi/10.1002/sam.11372. doi:10.1002/SAM.11372.

[6] R. Boselli, M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, M. Viviani, WoLMIS: a labor market intelligence system for classifying web job vacancies, Journal of Intelligent Information Systems 51 (2018) 477–502. URL: https://link.springer.com/article/10.1007/s10844-017-0488-x. doi:10.1007/S10844-017-0488-X/METRICS.

[7] I. Feinerer, K. Hornik, D. Meyer, Text Mining Infrastructure in R, Journal of Statistical Software 25 (2008) 1–54. URL: https://www.jstatsoft.org/index.php/jss/article/view/v025i05. doi:10.18637/JSS.V025.I05.

[8] G. Karypis, CLUTO - A Clustering Toolkit (2002). URL: http://conservancy.umn.edu/handle/11299/215521.

[9] G. Csárdi, T. Nepusz, The igraph software package for complex network research (2006).