Big Corpus of social media chatter - Big challenge for marketing scientists. The case of a 2.7 billion tweets Covid19 dataset.

Michel CALCIU
Université Lille, LUMEN,
mihai.calciu@univ-lille.fr

Jean-Louis MOULINS
Aix Marseille Université,
CRETLOG, jean-louis.
moulins@univ-amu.fr

Francis SALERNO
Université Lille, LUMEN,
francis.salerno@univ-
lille.fr

Abstract:

This research addresses the technological challenges encountered by marketing scientists when constructing and analyzing a large corpus of social media content. The corpus comprises 2.7 billion COVID-19-related tweets, covering the critical period of the pandemic and global crisis from January 2020 to February 2023, after which access conditions to Twitter/X were disruptively altered by Elon Musk. The primary challenge for marketing scientists lies in integrating computer cluster and cloud technologies to ensure scalability, flexibility, and computational efficiency in the management of Big Data and AI applications.
We demonstrate that building such a corpus constitutes fundamentally an ETL (Extraction, Transformation, Loading) process. The analysis of this corpus requires familiarity with Big Data technologies and concepts, such as MapReduce, to enable model computation at scale. We argue that the adoption of Big Data technologies through cluster and cloud computing has facilitated the development of deep learning approaches using neural networks, resulting in the emergence of large models, including LLMs (Large Language Models), which have largely supplanted earlier, smaller, and more traditional NLP (Natural Language Processing) techniques. Consequently, marketing scientists should prioritize LLMs over smaller statistical language models when analyzing social media data.
Several solutions we adapted for sentiment analysis on this corpus are subsequently introduced and discussed.

Introduction

Twitter/X is a microblogging platform that allows users to express their feelings, emotions, and opinions through short messages, enabling real-time circulation of information on a global scale. This wealth of information has provided significant opportunities for research. Compared to other social media platforms, Twitter's relatively straightforward access via the Twitter API made it, until recently, a primary resource for researchers. However, since February 2023, the Twitter API is no longer freely available following management changes after the company's acquisition by Elon Musk.

In this paper, we examine the value of a large corpus of tweets as a research resource for marketing scientists, covering the main period of a rather unique pandemic and worldwide crisis, in our case the Covid19 phenomenon. During this time, Twitter usage increased substantially, establishing the platform as a central venue for public discussion and dissemination of pandemic-related information and opinions.

Leveraging a massive collection of over 2.7 billion tweets gathered during more than three years of the COVID-19 pandemic presents significant challenges for marketing scientists. These challenges stem from the sheer volume of data to be collected, the computational requirements and acceleration technologies necessary to process such large-scale datasets, and the transition from traditional NLP methods and small machine learning models to increasingly large LLMs based on AI.

The remainder of this paper is organized as follows. First, we demonstrate that constructing this large corpus constitutes fundamentally an ETL (Extraction, Transformation, Loading) process. Next, we discuss how analyzing the corpus requires expertise in cluster and cloud computing technologies, as well as concepts such as MapReduce, to enable model computation at scale. We emphasize the importance of high-performance computing (HPC) resources available at the university level. Finally, we present our conclusions and propose directions for future research.

Constructing a big corpus - an ETL exercise

Our COVID-19 corpus comprises over 2.7 billion tweets obtained from a web repository covering the main period of the pandemic (January 2020 to February 2023). In accordance with Twitter's Terms of Service, only the tweet identifiers have been published (see Chen et al., 2020, 2025).

Constructing this corpus employed an ETL (Extraction, Transformation, Loading) approach, requiring data engineering expertise and data management tools (in this study, shell scripts and Talend). ETL is a standard process involving the collection of data from source systems, its transformation into a useful format, and its storage in a target repository, such as a database, data warehouse, or data lake. In our case, the process begins with a **data-sourcing** exercise, extracting tweet identifiers from the aforementioned external web repository. This repository contains 26,914 hourly text files covering 1,121.4 days—equivalent to three years and 26 days—from 21 January 2020 to 17 February 2023.

The frequency of tweets per hour varied considerably over the course of the pandemic. During the first two years, the period likely posing the greatest challenge, the average volume was approximately 125,000 tweets per hour, decreasing to around 45,000 tweets per hour in 2022 and early 2023.

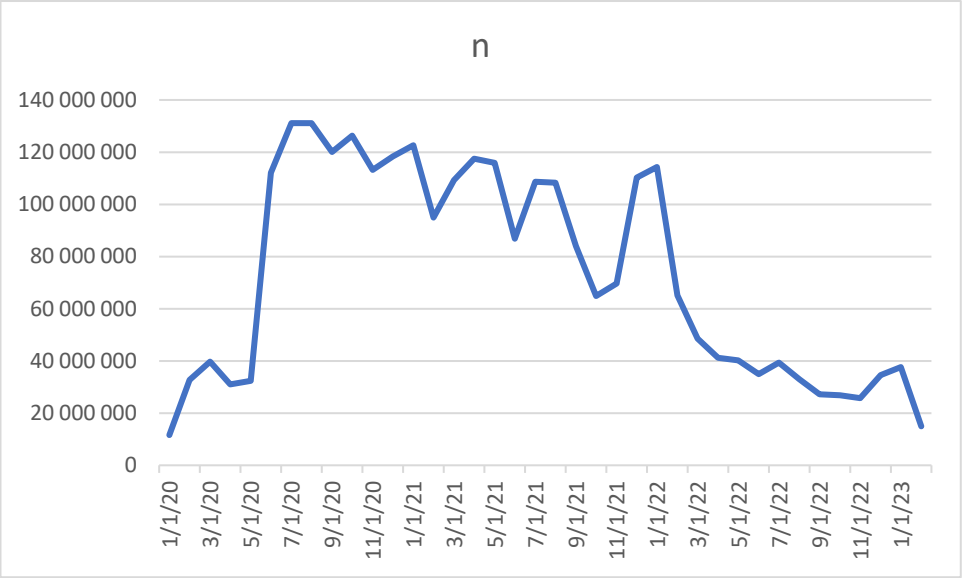

Figure 1 – Monthly Tweets in our Covid19 dataset

Figure 1 illustrates the monthly evolution of tweets over the specified period. From January to February 2020, pre-outbreak awareness emerged following the first Wuhan alert (30 December 2019) and the WHO's declaration of a global emergency (30 January 2020). Between March and May 2020, conversation on Twitter surged rapidly, reaching a peak of approximately 150,000–200,000 tweets per hour, driven by global lockdowns (e.g., Italy, 9 March 2020) and key WHO announcements (11 March 2020). From June to December 2020, activity oscillated and gradually declined as the novelty of the pandemic diminished. In 2021, a second growth phase occurred, fueled by renewed attention surrounding vaccine developments (e.g., Pfizer 90% efficacy, 9 November 2020) and subsequent spikes during the Delta (11 May 2021) and Omicron (26 November 2021) waves.

The "dehydrated" dataset can be accessed by cloning the corresponding GitHub repository (Chen et al., 2025). Due to the substantial size of the collection (~73 GB), this process may require several hours, depending on the speed of the user's internet connection.

The **extraction phase** involved hydrating tweets. Instructions for using the Twitter API with Twarc are provided by Chen et al. (2025). A Twitter developer account (https://developer.twitter.com/en/apply-for-access) was required, as access keys were used in the execution of the *twarc configure* command.

Twarc enabled the collection and hydration of tweet IDs, allowing the content of each tweet to be retrieved from its unique identifier and archived in Twitter JSON (JavaScript Object Notation) format. Additionally, Twarc managed the Twitter API rate limits, which were imposed to regulate the pace of data extraction.

On average, hydrating an hourly tweet ID text file containing approximately 130,000 tweet IDs required around 30 minutes, corresponding to a Twitter-controlled extraction rate of roughly 5,000–6,000 tweets per minute. Consequently, hydrating approximately 700 tweet ID files for a single month would take an estimated 370 hours, or about 15 days, assuming no technical interruptions. In practical terms, the extraction of our three-year COVID-19 tweet collection required approximately one and a half years to complete.

Although the extraction phase, involving tweet hydration, was the most time-consuming, the **transformation** phase entailed converting .gz archive files into *.jsonl*, a semi-structured file format specific to tweets using JSON, and subsequently into tabular *.csv* files. While .csv files retain only 37 fields per record, *.jsonl* files typically contain 80–250 fields per tweet object in

COVID-19 datasets, preserving much more detailed information. To support future research, the complete corpus was maintained in *.jsonl* format, occupying over 6 terabytes of storage on the university cluster's virtual disk, with additional copies stored on offline storage systems. For initial operations, we also selected 1.4 billion English-language tweets in .csv format and extracted random samples at rates of 1/100, 1/1000, and 1/10000 for analysis.

The final phase of this ETL process is **loading** the data into a database, which may be a structured SQL (Structured Query Language) database such as MySQL or a semi-structured NoSQL (Not Only SQL) database such as MongoDB. Databases offer the advantage of providing persistent storage while enabling flexible queries over large volumes of data.

A JSON file can be imported directly into a MongoDB database collection using a command like:

mongoimport --jsonArray --db corona --collection covid --drop --file ./coronavirus-tweet-id-2020-01-22-10.json

Importing a .csv file into a database table requires additional preparation which involves several steps:

1. Prepare and execute the SQL command that creates the database
2. Generate a draft of the SQL command to create the table using the field names from the .csv file headers;
3. add details about the field types.
4. execute the generated file to create an empty database table
5. After setting up the structure of a table that matches the .csv file, the latter can be imported and populate its records.

Although no acceleration methods were employed during the extraction and construction of our corpus, several powerful approaches exist that could have been applied to expedite these operations. The most notable include:

- Using the Ray package that is easier for Python functions with automatic parallelization.
- Using Spark on a Spark cluster which is more difficult to orchestrate but allows massive scaling.

The potential of this COVID-19 tweet collection for research employing Machine Learning, and particularly Deep Learning (DL), is substantial. The dataset appears to be comprehensive, covering the main period of the pandemic and overcoming current restrictions on access to Twitter data. This collection represents a valuable "data mine" capable of supporting a wide range of future research topics in marketing and management, including health and pharmaceutical marketing (Zaynab, 2024) and disaster or crisis management (Balech et al., 2022), among others.

High Performance Computing solutions for the big corpus

A major challenge in handling datasets of this scale is the effective use of high-performance computing (HPC) facilities to accelerate computations. The use of supercomputers can be prohibitively expensive, and public cloud services hosted by privately owned data centers, such as Databricks, Amazon AWS, Microsoft Azure, or Google GCP, also entail significant costs. Under these circumstances, free HPC resources provided by university data centers represent a viable alternative. These HPC facilities can be accessed either as "share everything" hybrid clusters or as "share nothing" private cloud services based on OpenStack, an open-source cloud computing platform.

Building a "share nothing" cluster and a Big Data Analytics (BDA) Platform on a private cloud

While public clouds provide Big Data and AI processing through their proprietary and often costly **P**latform **a**s **a** **S**ervice (PaaS) solutions, free private OpenStack clouds are particularly well-suited for offering **I**nfrastructure **a**s **a** **S**ervice (IaaS). However, IaaS presents a significant challenge for non-expert data scientists, as they are required to orchestrate both the Big Data and AI infrastructure and platform themselves—a complex and time-consuming task. We have described and applied this approach in a previous study (Calciu et al., 2020).

For this project, we constructed a computer cluster consisting of six virtual machines (VMs) or nodes, in accordance with the quota allocated by the university's cloud administrator. Each VM was provisioned with the largest available configuration, comprising six cores and 60 GB of RAM, resulting in a total of 36 cores and 360 GB of RAM. The infrastructure also included a network and virtual disks providing persistent storage of 7 TB for the COVID-19 corpus.

On this "share nothing" computational infrastructure, we developed a Big Data Analytics (BDA) platform based on Apache Spark, which remains one of the most powerful and widely used BDA computation engines. Our approach built upon a solution proposed by Borisenko et al. (2016), incorporating Ansible to automate provisioning, configuration, and deployment (see Calciu et al., 2020). Ansible is an open-source, command-line IT automation tool written in Python. To our knowledge, this represents the only such BDA platform constructed at our university and also the first such academic marketing attempt in France. Authors in a special issue on Big Data in Marketing Science (Liu et al., 2016) note that academia lags behind industry in cloud analytics, and they recommend that future marketing research consider adopting cloud-based tools, such as Spark.

We employed our BDA platform to predict consumer ratings using the Amazon Reviews dataset (courtesy of McAuley et al., 2015; file size: 58.3 GB), which contains 82.68 million reviews after deduplication (originally 142.8 million) spanning May 1996 to July 2014. Lasso regression was applied in this context, as it is well-suited for situations with a large number of independent variables—in this case, unigrams. The method automatically selects the most relevant features while discarding less informative ones. Remarkably, fitting the model on this massive dataset required only approximately 20 minutes.

Since Apache Spark does not provide native integration with OpenStack in the same manner as it does for AWS (S3, EMR) or Azure (ADLS), maintaining a BDA platform on private or public OpenStack clouds is relatively challenging. This limitation partly explains the discontinuation of Sahara, a significant OpenStack project with similar objectives. Furthermore, industry trends have shifted away from deploying Spark on large virtual machines toward container-based architectures. In future research, we plan to adopt this trend; for the remainder of the present study, however, we relied on 'share everything' solutions.

Using share everything hybrid cluster

At this stage, and particularly for our large corpus, the most practical solution for non-expert academic marketing scientists is the use of a university "share everything" hybrid cluster. As noted, this type of cluster provides shared access to all resources. Computing resources, including nodes (physical computers), cores (CPUs), and available software modules, are accessed via the open-source SLURM (Simple Linux Utility for Resource Management) software. SLURM requests the number of nodes and cores to be allocated and launches jobs—such as R or Python programs—that perform computations on large datasets concurrently across multiple nodes and multiple cores per node, leveraging both node-level and core-level parallelism.

Accelerating small calculations on big data

Lexicon-based sentiment annotation approaches determine the polarity (positive, negative, or neutral) of a text by evaluating the individual polarity of the words it contains (Cambria, 2016). Although these annotation processes typically involve relatively small computations, they can become computationally intensive when applied to datasets comprising millions of records, in our case tweets, potentially requiring hours or even days when executed on a single personal computer or cluster node.

To address this challenge, we developed a method called GRAPPA (GeneRal Approach for Parallel Processing Annotations) and applied it to annotate tweets (Calciu et al., 2021). This approach leverages R's *parallel* and *rslurm* packages (R Core Team, 2020). The *parallel* package enables multi-core parallelism, while the *rslurm* package facilitates cluster-level parallelism using SLURM.

Significant speed improvements were achieved on a sample of 1,401,244 tweets collected during the first 20 days of the COVID-19 lockdown in France. Computations that previously required six hours on a single node were reduced to eight minutes when executed on 16 nodes with two cores each (32 cores in total). In general, computation time scales approximately linearly with the number of cores utilized.

Considering a larger sample of approximately 14 million tweets, extracted from 1.4 billion English-language tweets in our corpus, annotation on a single core would require roughly 60 hours. Using 10 cores, the computation time decreases to approximately 19 hours; 100 cores reduce it to 1.9 hours, and 1,000 cores reduce it to approximately 11 minutes.

However, there are inherent limitations in a "share everything" environment. The university hybrid cluster comprises 6,664 cores, and the more users share these resources, the longer the computation time required. Additionally, the maximum wall-time constraints increase with the number of cores requested: for fewer than 48 cores, the maximum wall-time is 16 days, whereas for 768 cores, it is reduced to one day. Such constraints do not exist in a "share nothing" environment, where resources are dedicated to a specific quota of virtual machines.

Accelerating big model calculations on big data

Accelerating computations on large datasets with large models, such as LLMs, introduces at least two additional challenges. The first is the replacement of statistical software systems, such as R, with more suitable ecosystems for LLMs, such as Python. The second is the adoption of memory-efficient parallelism approaches, exemplified by Python's Ray module (Moritz et al., 2018). In multi-core parallelism, the available cores share the same memory within a node, whereas in cluster-level parallelism, each node possesses its own memory, the size of which can be controlled using SLURM.

Our transformer-based model, RoBERTa, was fine-tuned on Twitter data and specialized in sentiment analysis. It is a high-performing large model with a relatively manageable memory footprint of approximately 2–3 GB.

To achieve both node-level and core-level parallelism, we employed the hierarchical task scheduling capabilities of the Ray module. In the SPLIT phase, the array of texts (tweets) is divided into a number of chunks corresponding to the available nodes (machines) in the cluster. Each node-level chunk is further subdivided into sub-chunks according to the number of cores available on that node. These node chunks and core sub-chunks are stored as lists of split text arrays. During the MAP phase, the annotation function *model.sentiment(tweet)* is applied to each sub-chunk, producing sentiment predictions or inferences.

Inference on a sample of 1.5 million tweets using two nodes without core-level parallelism required nearly one day (approximately 18 hours). When executed on 32 nodes with 24 CPUs each, the same task was completed in under 11 minutes per node on average. Utilizing a "share everything" cluster in this manner enables a substantial reduction in inference time for large

models applied to large corpora. For instance, processing a sample of 15 million tweets would take nearly ten days on a single machine with one CPU, whereas the 32-node × 24-core configuration reduces the computation time to slightly more than one hour. Such efficiency gains make the application of deep learning models, including LLMs, considerably more feasible for academic marketing researchers.

Discussion and conclusion

This research primarily investigates strategies for accelerating computations on large datasets and large models (LLMs) using high-performance computing (HPC) resources in university cloud and cluster environments. The study demonstrates that substantial time savings can be achieved through core-level and cluster-level parallelism. However, two Hamletian dilemmas arise:

1 - “To big, or not to big, that is the question: whether ‘tis nobler” for marketing scientists to invest in big data technologies or not. One aspect is clear: the democratization of Big Data, facilitated by Google’s introduction of the MapReduce approach (Dean & Ghemawat, 2004), enabled the massive parallelization of computations that were previously limited to single machines. This development has contributed significantly to the rapid advancement of AI and the proliferation of LLMs.

For many marketing scientists, it remains uncertain whether large-scale data is essential, particularly given that sampling—especially opinion sampling—is a foundational methodological practice in marketing research. While Big Data can often be substituted with representative samples, large models cannot. Consequently, the application of reasonably large transformer-based LLMs to sufficiently large datasets encourages the adoption of these computational approaches. This, in turn, highlights the need for marketing scientists to acquire additional system engineering skills, as previously discussed, to efficiently leverage computer clusters and cloud infrastructures for accelerated computation.

2 - “Too big, or not too big” is another critical question, referring to the increasing scale of data and models, particularly LLMs, and whether they can still be managed within conventional academic research settings. Although today’s LLMs share a common architecture—large transformer-style models pre-trained at massive scale using next-word prediction—they are trending toward industrial scale, with parameter counts approaching the trillions.

We are currently experiencing a dynamic and, at times, confusing period in society, in science broadly, and in marketing science specifically. This era of rapid change, driven by the rise of Deep Learning AI and facilitated by Big Data technologies, demonstrates how computationally efficient and scalable methods can generate complex systems that increasingly surpass human expertise. Marketing scientists who wish to engage with or monitor these technological developments should prioritize familiarity with high-performance computing (HPC) infrastructures, including cloud and cluster-based technologies.

Reference

Balech S., Benavent C., Calciu M. & Monot. J. (2022) Le masque, figure polaire de la crise de la Covid-19 : une exploration par NLP du flux des conversations Twitter (février-mai 2020), *Marché et organisations*, 1/ 43, p. 151-187

Borisenko, O., Pastukhov R., Kuznetsov S. (2016), Deploying Apache Spark virtual clusters in cloud environments using orchestration technologies, *Trudy ISP RAN/Proc. ISP RAS*, 28, 6,111–120, DOI: 10.15514/ISPRAS-2016-28(6)-8

Chen E., Lerman K., Ferrara E. (2020), Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set, *JMIR Public Health Surveillance,* 6, 2, 1-9

Chen E, Lerman K, Ferrara E. (2025), COVID-19-TweetIDs, https://github.com/echen102/COVID-19-TweetIDs, [accessed 2025-10-05]

Calciu M, Moulins J-L. & Salerno F. (2020), Marketing Knowledge Discovery and Big Data Analytics. Towards reducing technological entry barriers for marketing scientists, *19th International Marketing Trends Conference*, Paris, January, 16-18

Calciu M, Moulins J-L. & Salerno F. (2021), The GRAPPA method for accelerating annotations on Big consumer opinion datasets. Applications to sentiment modeling on COVID19 Lockdown Tweets and Amazon Reviews. *20th International Marketing Trends Conference*, Venice, January, 14-16

Cambria A. (2016), Affective computing and sentiment analysis, *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

Dean J. & Ghemawat S. (2004), MapReduce: Simplified Data Processing on Large Clusters, *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, December.

Liu, X. Singh, P.V. & Srinivasan, K. (2016), A Structured Analysis of Unstructured Big Data by Leveraging Cloud Computing, *Marketing Science*, 35, 3 (May/Jun), 363-388.

McAuley, J., Pandey, R. & Leskovec J (2015) Inferring networks of substitutable and complementary products, *KDD '15 Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Moritz P., Nishihara R, Wang S., Tumanov A., Liaw R., Liang E., Eliahu M., Yang Z., Paul W., Jordan M.I., Stoica I. (2018) Ray: A Distributed Framework for Emerging AI Applications, *13th USENIX Symposium on Operating Systems Design and Implementation*, Carlsbad, 561-577, ISBN, 978-1-939133-08-3, https://www.usenix.org/conference/osdi18/presentation/moritz

R Core Team (2023) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Viena, https://www.R-project.org/

Zaynab Z. (2024) " Big pharma's electronic word of mouth during Covid-19", 23rd International Marketing Trends Conference, Venice, January 18-20