SoDa-TAP : A Data Platform for Social Media Analysis

by

Candelario Alfonso Gutiérrez Gutiérrez

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Social media platforms are online public venues where conversations about a wide range of public interests take place. Users can interact within a social platform in two ways: (i) they can post an opinion, talk about an event, or share a personal status, optionally accompanied with a video, an image, url(s), and/or hashtag(s); or (ii) they can like, share, reply and/or quote another user's post. Social platforms are known to generate high volume of interactions: a large number of users are frequently participating in online conversations, and share their ideas or thoughts on issues. Private and public organizations have shown interest in the data that is generated from these interactions with the purpose of monitoring topics, events, brands, and/or persons of interest. This data then allows them to make informed, event-driven decisions. In order to achieve this, there is a necessity for Social Media Analytics systems that facilitate the collection, processing, storage and extraction of insights from social media data.

In this thesis, we present *SoDa-TAP* (Social Data - Toolkit Analysis Platform), an automated, scalable, and extensible data platform that offers three key functionalities: (a) Data Workflow: data extraction of posts' elements and metadata, calculation of secondary metrics, and multiple visualizations for data exploration; (b) Engagement Calculation: analysis and calculation of multiple indicators of influence from the posts' data; and (c) Statistics Toolkit: a toolkit for descriptive statistical analysis and information visualization. We demonstrate the platform's potential in two ways. First we describe the software architecture and how the interconnectivity of each component supports fast and scalable deployment. The second is the presentation of two case studies that we performed, to show the ability to measure influence using official Twitter accounts from the University of Alberta, and its adaptability to provide insights about conversations and content engagement around energy, during the “Energy East Pipeline” timeline.

Preface

This thesis is an original work by Candelario Alfonso Gutiérrez Gutiérrez. Segments from this thesis has been published in the following literature:

- **C. A. G. Gutierrez**, A. Whittaker, K. M. Patenio, J. Gehman, L. M. Lefsrud, D. Barbosa, and E. Stroulia, "Analyzing and visualizing Twitter conversations," in Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering, 2021, pp. 4–13.

- **C. A. G. Gutierrez**, J. Gehman, L. M. Lefsrud, D. Barbosa, and E. Stroulia. (in press) "Energy to Contest? Emotional and multimodal contestation of energy markets," European Group for Organizational Studies (EGOS), Vienna, July, 2022.

Acknowledgements

I would like to thank my supervisor, Dr. Eleni Stroulia, for her support, encouragement and patience. Your expertise, constructive feedback and guidance steered me to the right direction throughout this research. I have no words to express my gratitude towards the trust that you put in me.

I would like to thank Professors Lianne Lefsrud, Joel Gehman, and Denilson Barbosa, and at the same time the students Katherine Mae Patenio and Andrea Whittaker for their valuable time, advice and help during the design, development and demonstration of multiple components that are presented in this thesis.

I want to thank Lianne Lefsrud and Ebrahim Bagheri for serving on my defense committee.

A special thanks to Kalvin Eng, Mashrura Tasnim, Victor Fernandez Cervantes and Christoph Sydora, for fostering a really collaborative lab.

To conclude, I want to thank my lovely family for their unconditional lifetime support and encouragement, and friends that were there who listened and helped me along this journey.

Thank you.

Candelario Alfonso Gutiérrez Gutiérrez

Table of Contents

List of Tables

List of Figures

Abbreviations

ABSA Aspect-Based Sentiment Analysis.

ANOVA Analysis of Variance.

API Application Programming Interface.

BI Business Intelligence.

CI Competitive Intelligence.

CI/CD Continuous Integration/Continuous Deployment.

ML Machine Learning.

NLP Natural Language Processing.

SMA Social Media Analytics.

SOCMINT Social Media Intelligence.

Listings

Chapter 1

Introduction

The term “Social Media Intelligence (SOCMINT)” refers to a broad research area that integrates data analytics and data mining techniques applied to content published and consumed through social channels [1]. This area has gained popularity as a subject of study because of its data source connection with the real world. People use social platforms to share what they consider relevant to their personal lives, and through this type of technology, a number of very interesting use cases have emerged, ranging from following an individual’s posts to understand their beliefs and behaviors, to analyzing a community’s discussions to measure the impact of public events and the public’s opinions on key matters. Social platform interactions are worthy of study, as they revolve around multiple actions that have an impact on society [2]. Some interesting use cases of social-platform analysis include the following [2]:

- Natural disasters alert, run by official organizations and community involvement to notify the public about safety measures;
- Traffic notifications to share different events that are happening on the road such as traffic jams, accidents, traffic regulators, etc;
- Opinions that people share about products, services and brands, to let others know what to expect when buying; and
- Public opinion from different points of view on current events, such as regula-

tions, elections, stock market, projects' development, etc.

The above reasons have motivated the development of a variety of algorithms and tools, to comprehend different aspects of such interactions during these events, such as topic discovery, user influence, social network analysis, user identification and clustering, message heuristics and persuasion, and many more [3]. In parallel with the research advances, the task of social-media intelligence evolves. Access to a smartphone or to a computer is relatively easy, and, as a result, people feel the necessity to share via their socials their day-to-day activities and events. Therefore, the social-platform data stream continuously increases in volume and velocity, which is the reason why new software and technologies, with big data capabilities are necessary, for an automatic and extensible analysis, with the benefit of leading insights' discoveries and the extraction at scale from different sources.

1.1 Motivation and Background

The landscape of social media platforms is broad and includes social networking services, blogs, social bookmarking sites, content sharing sites, and opinion sharing sites [4]. Each of these platforms serves a specific use case. For example, Facebook and LinkedIn are used to create online communities and personal/professional networks; Tumblr and Twitter are micro-blog services to frequently share short messages; Reddit is used to share bookmarks about a variety of content; YouTube and Instagram are used to share multimedia content; Glassdoor, Yelp and TripAdvisor are used to rate and leave opinions about different experiences. According to Kaplan and Haenlein [5], social media platforms can be characterized in terms of two dimensions. "Influence" involves the users' virtual representation of their personas, through which they try to control how they are seen and perceived by others, online. "Media richness" involves the amount of information that each persona shares at different times. Each social media platform offers different options of influence and media richness, based on the types of interactions they afford their users, such as, for example, “like”, “share”, “comment”, etc. Such mechanics can make the composition of a post simple or it can be converted into something really complex if the platform allows to incorporate more than just text, like in the case of adding media to the message.

The variety of use cases, social-platform interactions, and user-provided data and metadata are well explored. Many algorithms have been developed to understand online user behaviors in different contexts: politics, public opinion, risk alerts, public policy development, etc. Today, this type of data is mostly seen as an opportunity by interested companies to better understand their customers and to evaluate their brand reputation online [6]. Furthermore, organizations and governments are using social-media analytics to understand their community and to get to know how they feel about multiple societal aspects [7].

These reasons have motivated our work on *SoDa-TAP* , an extensible and scalable tool that can transform the voluminous users’ online interactions into insights, to support event-driven decision making, through the leverage of different tools and systems combined with statistical capabilities. Batrinca and Treleaven [8] propose that there are three important elements that should be considered when developing such a tool: (i) data, the source of the social data that should be of open access; (ii) analytics, the actions performed over the data, i.e., graphical plots, custom analyses and applications that allow its access for third-parties; (iii) facilities, infrastructure that is used to host huge volumes of data, and to deploy the analytics component, powered by systems and tools that make use of computing resources that are exploited to process the data. *SoDa-TAP* integrates a set of tools that enable the analysis of this type of data that are either developed by a social-platform organization or by third-party companies that hold an agreement with the social platforms, such as Facebook audience [9] and Instagram insights [10], Twitter analytics [11], BuzzSumo [12], Falcon.io [13], Hootsuite [14], Brandwatch [15] and others [16–19]. The main characteristics of these “official” platforms are that they offer a free set of metrics alongside a dashboard, to visualize detailed information about an account's actions and content engagement. The third-party analysis platforms offer similar functionalities, and can additionally support the connection of multiple accounts in a single platform. They contribute a deeper analysis by giving its users the capabilities of (a) creating custom dashboards with a set of available charts; (b) tracking conversations based on custom keywords; (c) providing insights about a topic and trend discovery based on popular hashtags. At the same time, these platforms typically experience the following limitations: (i) there are a limited set of operations that you can perform; (ii) some of them require a paid subscription; (iii) you need to manually adjust the systems' configuration. There are several key elements of complexity in developing a social-platform analytics system:

1. Big-data software frameworks are complex and difficult to work with. Usually, these tools need manual intervention to work with different workloads and in specific environments. This increases the learning curve time for the developer(s).

2. Recent research about conversations' context relies mostly on trained models. This adds more complexity to the platform and makes data processing take more time due to computing resources.

3. The majority of platforms are only limited to analyze text and posts' metadata. Pictures, posts' elements, multi-modal and complexity, URL(s), and statistical analyses are not fully contemplated.

4. Already developed platforms, applications and frameworks like "Communalytic" [20], "Netlytic" [21], "DiscoverText" [22], "Social Bearing" [23], "SocioViz" [24] and others [25–31], are not adaptable or modular enough in order to extend their functionality for custom use cases or analyses.

1.2 Contributions

We have developed a modular and scalable platform called *SoDa-TAP* (Social Data - Toolkit Analysis Platform) with the capability to analyse Twitter data. Our platform has been designed and implemented following the micro-service and containerized architecture in order to achieve the flexibility of module connection and/or replacement, to positively improve the data flow and complement the analysis capabilities.

This thesis makes three contributions to the field:

1. **The *SoDa-TAP* System** delivers the following key functionalities.

   A set of clients to fetch Twitter posts. We integrated a set of Twitter clients to support different data search criteria, such as time delimited posts, users' timeline and/or keywords to find in the post.

   The data processing and analysis engine. We designed a set of data processing and analysis pipelines that are able to recognize and work over social media data. In the case of the data processing, it was developed under the paradigm of parallel computing, with the flexibility to add a streaming or offline source, and for the analysis of text, it was developed following a dictionary lexicon-based approach.

   A set of Application Programming Interface (API)s to enable third party applications to consult the database. We developed a set of endpoints that listen to requests to let external applications read and write data into the database.

   An API for image analysis. We developed an endpoint that performs and aggregates inference results such as color scheme, object detection, image classification, sentiment, Optical Character Recognition (OCR), and face detection.

   A set of visualizations to help describe the processed data. We developed a set of custom visualizations to help with the interpretation and insights discovery of the processed data.

A statistical analysis component that allows to perform Analysis of Variance (ANOVA). We developed a component that allows to perform one-way, two-way and n-way ANOVA over custom defined variables. This component is also in charge to show a box plot per analysis.

2. **An automated deployment and processing pipeline for the platform.** *SoDa-TAP* has the capability to be deployed without human interactions, thanks to a custom defined configuration file, with the purpose to reduce deployment complexity. This characteristic allows our platform to be a perfect fit for a Continuous Integration/Continuous Deployment (CI/CD) pipeline, giving to the developers or researchers the opportunity to set up all the infrastructure in less time.

3. **Custom analyses** to demonstrate the effectiveness of *SoDa-TAP* in extracting, processing, analyzing and presenting insights from social data. These studies also demonstrate the usefulness of *SoDa-TAP* in examining key hypotheses on how opinions are shared and propagated through social platforms. Our insights into these hypotheses can help in the decision-making process within an organization or for research purposes.

1.3 Thesis Outline

The rest of this thesis is organized as follows. Chapter 2 reviews the background and related research on social media platforms. Chapter 3 describes the design, components and implementation of *SoDa-TAP* . Chapter 4 describes our experience with the application of the platform. And finally, chapter 5 concludes with a summary of our work, showcasing our key findings, and discussing our future plans for further *SoDa-TAP* features' development.

Chapter 2

Background and Related Research

Recently, there has been an increase in Social Media Analytics (SMA) applications and tools because of the high volume of online interactions taking place on social media platforms [7]. These interactions are linked to people's real-life activities all over the world. Health care, education, crisis management and business are some of the impacted fields that people's activities are related to the most, when online [32, 33]. Because of the nature of the data generated from these online interactions, a variety of research fields have shown interest in this type of data and have made it their object of study thanks to the ability to access this data through different services and APIs [8]. SMA applications are developed with specific use cases and constraints that need the selection of the proper algorithms and methodologies, depending on the research field [34]. Hence, SMA tools and frameworks have adopted and integrated the necessary workflow to collect, process and visualize this type of data [35], but most of these solutions are limited in their extensibility to adapt to a wider set of use cases [36].

SMA applications take components from Competitive Intelligence (CI) to increase the degree of insight discovery, give more context for the decision-making, and to help take faster event-driven decisions by unifying the data workflow and tools that are used to analyze data [37]. SMA applications can be used for a variety of quantitative and qualitative studies, for example: influence analysis, to calculate user centrality, message diffusion and ranking scores [38–42]; people's opinion and interactions sentiment analysis, to study conversations' discussion, the relationship between consumers and brands or understand emotions from online interactions [43–45]; and content engagement analysis, to study visual aspects and composition of a post and its engagement [46–49].

To be able to allow SMA applications adapt to different studies, a diversity of algorithms, methods and tools are required to cover more ground of the types of analyses that can be done [50]. Hence, "extensibility" is an essential architectural design that these systems should contain, to be able to include a variety of data sources, and visual components, to adapt to the state-of-the-art techniques for SMA.

2.1 Social Media Analytics Techniques

As social platforms keep gaining popularity, the volume of the data increases. The openness that these platforms offer to their users and what makes them compose messages and interact with posts, have been previously studied and linked to six "factors" [51]: (1) users like the anonymity that a platform offers to share ideas about any topic; (2) users' interaction depends on internal preferences to feed their own necessity of interaction, whether their need is to just share, to support other users or to generate controversy, through a message; (3) users try to find posts, groups or content that align with their profile or topics of interest; (4) online actions might be easier than performing an offline version of them; (5) online presence of politics has the power to attract users and leads to information exchange; and (6) users' profile, values, character, behaviour and other personal traits align to their activity online. Depending on the purpose of the study, the desired analyses, and the hypotheses to be evaluated, a variety of SMA algorithms and methods are being developed. The following are the respective description of some of the most popular [52–56]:

- **Sentiment:** Also known as opinion mining. Type of analysis to extract and quantify emotional related information. Normally categorized as positive, negative or neutral. This type of analysis is usually done over text and images.

- **Social Network:** Type of analysis to extract relationships between groups, people, objects or topics that share a converging point. This technique is more frequently used over social platforms data to understand people's online connection and community and to study those connections' structure characteristics.

- **Content:** Type of analysis to extract relationships, meaning, context and form within a collection of objects of study, usually texts or images, to analyze visual/textual patterns of expression related to a topic, product or event.

- **Statistics:** Type of analysis to interpret quantitative or qualitative results through the analysis of their numerical relationships, independently or as a group. This type of analysis is usually performed to understand the shape of the processed data results and to perform correlations or comparisons against custom defined variables.

These techniques have widely been adopted to mainly demonstrate case studies or to support hypotheses for a variety of applications. Individually, they offer approaches and algorithms that allow more granular analysis of the data. To mention some of them, **document, sentence, and aspect-based sentiment analysis** can be applied to recognize the emotion of documents, sentences, phrases or words, respectively [57, 58]. **Influence, centrality, and cluster analysis** can evaluate a social network connectivity and give a descriptive measure of users, groups or topics [59]. **Part-Of-Speech tagging, n-grams, bag of words and dictionary-based keywords frequency** can be used to evaluate textual context and composition [60–62]. **Correlations, regressions, comparisons and relationships** between groups can be applied over processed data results, in order to solve specific research questions that need numerical evidence to support a hypothesis [63].

As mentioned above, SMA applications development really depends on what the researcher, business or organization is seeking to infer for from its data. The techniques mentioned above can help to resolve part or all of the questions related to what they are interested in knowing. Nevertheless, the end-to-end process to evaluate this data, from the source to the visual results, can be quite challenging. SMA applications require a rigorous plan supported by tools, technologies, and frameworks to design and build an integral solution for data analysis.

2.2 Social Media Analytics Framework for Data Analysis

We came across different frameworks during our discovery phase of the SMA area. In our research, we found out that there are vast number of applications, methods and approaches that have been developed and defined based on common factors for the processing of data. The purpose of these frameworks is to help solve technical challenges related to CI and big data, mainly in the analysis component [6, 56].

Fan and Gordon [64] introduce the *CUP (Capture, Understand, and Present)* framework. They describe three key stages that social media analysis should follow: (i) *"Capture"*, select and collect objectively selected data from different online sources, services, and platforms; (ii) *"Understand"*, apply approaches and techniques from SMA to understand the data; and (iii) *"Present"*, apply visualizations and statistical analysis to support and convey a resulting message from the processed data.

Stieglitz, Mirbabaie, Ross, and Neuberger [65] expanded on a framework that was originally introduced for a political analysis perspective [66]. Their framework consists of four stages: *"discovery, tracking, preparation, and analysis"*. The analysis steps begin by first making an exhaustive look up of hashtags, keywords or words related to the topic of interest. Next, these set of keywords and words need to be searched in social media platforms, which will act as filters. Once the data is collected, a set of standardized steps need to be performed as an intermediary step called "preprocessing" that will serve to discard noisy data. To finalize the process, a set of techniques such as sentiment, content or social network analysis need to be performed over the preprocessed data to get insights out of it.

The framework proposed by Srivastava, Kumar and Narain [52], which inspired the data workflow in our system, elaborates on the previously mentioned framework to follow five stages: *"acquisition, pre-processing, representation, analysis, and presentation"*. The description of their data workflow starts by gathering data from social platforms through their respective API. Once all data is collected, a preprocessing pipeline removes unnecessary data/words and converts text into their base form, through stemming and lemmatization. All preprocessed data is later analyzed through a set of different pipelines to apply SMA techniques of their choice and to conclude the analyses, the results obtained from individual techniques are exposed through a set of visualizations.

As it is important to have a systematically way to work with social media data, the appropriate environment is also necessary to be met to handle specific tasks from the data workflow, which can be achieved through a rigorous selection of tools and technologies [6].

2.3 Technologies for Social Media Analytics Applications

Nowadays with the fast development of technology, there are many options available that can make a reality the implementation of any data processing framework. This variety of options can help with the development of SMA applications no matter the conditions for data processing, whether it is in an offline or in a streaming fashion [67, 68]. Based on the literature's data workflow stages and data processing constraints [6, 52, 56, 64–66], the following are some of the technologies that we discovered from our research in the area of data ingestion, processing, storage, and visualization:

- **Data Ingestion.** Component that is used to transfer or transport data. Apache Kafka [69], Apache Flume [70], Apache NiFi [71], and Apache Pulsar [72] are distributed and scalable solutions that can work in offline or streaming scenarios.

- **Data Processing.** Component that is used to apply SMA techniques, Extract Transform Load (ETL) operations or to manipulate the data. Apache Spark Structured Streaming [73], Apache Storm [74], Apache Samza [75], Google Dataflow [76], and Amazon Kinesis [77] can be used for real-time processing. Apache Spark [78] and Apache Flink [79] can be used for online/offline processing.

- **Data Storage.** Component that is used to store, retrieve and run queries over the processed data. Most of these technologies are distributed and from the NoSQL family. CrateDB [80], TimescaleDB [81], Azure Cosmos DB [82], Prometheus [83], CockroachDB [84], InfluxDB [85], Apache Pinot [86] are databases that can support offline and real-time use cases.

- **Data Visualization.** Component that is used to build charts and dashboards. Apache Superset [87], Tableau [88], Redash [89], and Metabase [90] are solutions ready to connect a data source and create dashboards based on a selection of available charts. ZingChart [91], Vega [92], D3 [93], Chart.js [94], and Google Charts [95] are web libraries to develop custom charts.

The integration of industry built and tested technologies, their complexity, adaptability, and orchestration are research topics that are worth to be explored from the software engineering perspective. At the same time, they open the door to the endless possibilities to create a complete SMA platform solution.

2.4 Social Media Analytics Platforms

As seen in Table 2.1, nine platforms are compared, including ours. These platforms support different SMA techniques to be applied over the data. A brief description of their functionalities and capabilities is shown below.

DiscoverText [22] is a web application that helps to work with text data through three stages: “collect, clean, and analyze”. Its core functionality is to provide an easy accessibility to data science tasks for collection, annotation, sampling, deduplication and clustering, filter, and sort data, extract uni, bi, and trigrams, and create reports of metadata metrics, such as top values, followers, following, and influence scores.

VosonSML+VosonDas [96] is a modular software package platform to build custom analyses through the R programming language. It is composed of two main libraries, vosonSML that allows to collect Twitter, Youtube, Reddit, and links from websites, extract sentiment and word frequency, and filter/sort all data; vosonDash that allows to interact with the data through social network analysis and create custom dashboards to visualize metrics data.

SOCRATES [97] is a web application platform composed of three blocks: (i) collect, a module to collect data from Twitter, The New York Times, Reddit, Facebook, Flickr, and Youtube; (ii) explore, a module that allows the exploration of the collected data as a table, histogram, scatterplot, and piechart; and (iii) analyze, a module to apply regression analysis over sentiment and word frequency of text.

Communalytic [20] is a web application platform that supports a variety of analyses over online conversations from social media data. It allows to import and collect Reddit, Twitter, Facebook, and Instagram data, to later be able to apply available SMA techniques, such as social network, sentiment, and toxicity analysis.

Netlytic [98] is a web application platform that allows to import Twitter, Facebook, Reddit, text, and RSS, that focuses on social network analysis to discover topics and do filtering based on keywords.

Table 2.1: Comparison between Platforms.

	Concept	**DiscoverText**	**VosonSML + VosonDash**	**SOCRATES**	**Communalytic**	**Netlytic**	**Mozdeh**	**Chorus**	**SocioViz**	**SoDa-TAP**
Data Source	Collect Datasets	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Import Datasets	✓	✓		✓	✓	✓		✓	✓
Data Presentation	Visualizations	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Sort/Filter Data	✓	✓	✓	✓	✓	✓	✓	✓	✓
Platform Development	Custom Deployment		✓	✓						✓
	APIs	✓								✓
	Extensibility		✓	✓						✓
SMA Technique	Sentiment Analysis	✓	✓	✓	✓	✓	✓	✓		✓
	Influence Metrics	✓			✓				✓	✓
	Dictionary Analysis	✓				✓	✓			✓
	Image Analysis									✓
	Statistical Analysis	✓		✓	✓		✓	✓		✓
	Network Analysis		✓		✓	✓	✓	✓	✓	
	Topic Analysis	✓				✓	✓	✓		✓

Mozdeh [99] is a desktop application platform for text analysis from Twitter, YouTube, Reddit, formatted, and unformatted data. It helps with the collection of the previously mentioned social platforms and it allows to run a variety of analyses: sort/filter data, topic, sentiment and influence extraction, time series navigation, and calculate word frequencies.

Chorus [28] is a desktop application platform composed of two components: (1)

"Tweetcatcher", a client to collect Twitter data based on keywords or through users' handle; and (2) "Tweetvis", a visualization window that allows to do data exploration to sort/filter data and a social network explorer to locate topics and tags of conversations.

SocioViz [24] is a web application platform capable of searching for words, handles, emojis from Twitter and Facebook. It allows to filter/sort data by dates, language, and specific keywords to visualize them in an area chart. Its core capability is the creation of a social network maps from users, hashtags, words, and emojis, to identify influence based on the previous mentioned components.

Each of these platforms have specific use cases that can help in the insight extraction from social media data, but as seen in Table 2.1, there are four key functionalities that differentiate *SoDa-TAP* from other platforms:

1. *SoDa-TAP* supports the curation of data sets in a variety of cases. Keywords, users or list of tweets can be used to create data-sets.

2. *SoDa-TAP* supports decision making through a variety of visualizations and data exploration through sort/filter functionality on a table.

3. *SoDa-TAP* can be extended to include new modules programmed for custom use cases, a set of API endpoints can be used whether to access the data for sampling or to use it as a source to serve visualizations, and it has the capability to automate its deployment as a local or in the cloud solution to reduce manual intervention and errors in computing environments.

4. *SoDa-TAP* includes SMA techniques to extract insights from text and media. Since it is modular and holds the extensibility feature, new algorithms and methods can be incorporated to expand the calculation of metrics.

The purpose of the previous comparison is to demonstrate the gaps and weak areas that have been noticed in the current SMA platforms. The described functionalities can be used to lead their future development, to develop more generic solutions that could simplify the replication and design of research studies, and to standardize the internal data workflow.

Chapter 3

Design and Implementation of *SoDa-TAP*

Batrinca and Treleaven [8] state that there are three important aspects to social media analysis: data, analytics, and facilities. Data can come from many different open sources, in an offline or streaming fashion. Analytics capabilities depend on the use case to be studied, through the application of different methods from Natural Language Processing (NLP), artificial intelligence, data mining and big data. Facilities involve the interconnection of many different tools and systems that make it possible to create end to end analysis pipelines. Other elements can be incorporated to extend such analysis capabilities and support the insight discovery phase, like the consumption of API endpoints that could contribute in specific analyses, or already developed Business Intelligence (BI) tools that allow to connect a database and run interactive queries for a specific data search. Reflecting on limitations, challenges and opportunities on related research and commercial systems for social media analysis [100–107], we have considered the following requirements in developing the *SoDa-TAP* toolkit:

1. It needs to be modular and extensible to integrate new components and use external systems. At the same time, new data sources, and Social Media Analytics (SMA) operations should be able to be easily developed and integrated. This opens the possibility to read and analyze other types of text such as news, documents, etc.

2. It should be able to analyze text and media. Social platforms' post(s) can be composed of only text, media or a combination.

3. Offline and online data analysis should be able to be performed within the same engine. By defining an architecture that allows to have the best of both worlds, it can enhance data discovery and make better insights based on historical and real-time data.

4. Depending on the type of the task, it should be able to perform it in multi-processing or multi-threaded fashion. Jobs that do not need to use a network connection should be performed with multiprocessing and if is the opposite, multi-threading.

5. Data storage access needs to have low latency and it needs to be easy to be plugged into third-party components.

6. Its infrastructure should allow it to scale up or down, depending on the computational resources available. Automatic components deployment and minimum manual intervention is an important characteristic to speed up development and integration.

7. A set of visualizations should be available to present the analyzed data and new graphs must be easy to be developed.

SoDa-TAP supports the deployment of data-processing pipelines to analyze social-platform posts. A pipeline may include the following processing components: (i) lexical analysis based on special-purpose curated dictionaries, such as dictionaries classifying words in terms of personal values, sentiment and humour; (ii) analyses to quantify the influence of posts and their authors; and (iii) image analysis based on deep neural networks to extract information about its color scheme and content.

The results of a *SoDa-TAP* pipeline can be visually explored through a rich set of visualizations. In addition, *SoDa-TAP* includes a Jupyter notebook[1] that allows users to investigate the data through statistical analyses, such as, for example, ANOVA to explore correlations between different post features.

An important aspect that was considered in *SoDa-TAP* was the decoupling of tasks that require external processing and requests. In this case, *SoDa-TAP* has incorporated two different functionalities: (a) image(s) download and (b) URL(s) expansion. Both of these functionalities are tasks that don't benefit from multiprocessing, since they rely on external resources. Usually, tasks that make use of the network are implemented using multi-threading to allow the creation of several tasks in different threads without worrying about them being blocked at any point, and without having to wait for the previous tasks to be finished which accelerates external resources consumption.

A *SoDa-TAP* pipeline can be deployed through the specification of its configuration, i.e., the included components and their required options, in a configuration file. Once a *SoDa-TAP* configuration has been specified, the corresponding *SoDa-TAP* pipeline can be automatically deployed, without the need of manual intervention.

3.1 The Software Architecture

As shown in Figure 3.1, *SoDa-TAP* consists of the following components: (i) Data Collection, (ii) Data Processing and Analysis, (iii) Data Storage, and (iv) Data Visualization.

3.1.1 Data Collection

The analysis pipeline starts with the collection of raw data from a social platform, in our case Twitter. Figure 3.2 shows what elements are currently extracted from a

[1] https://jupyter.org/

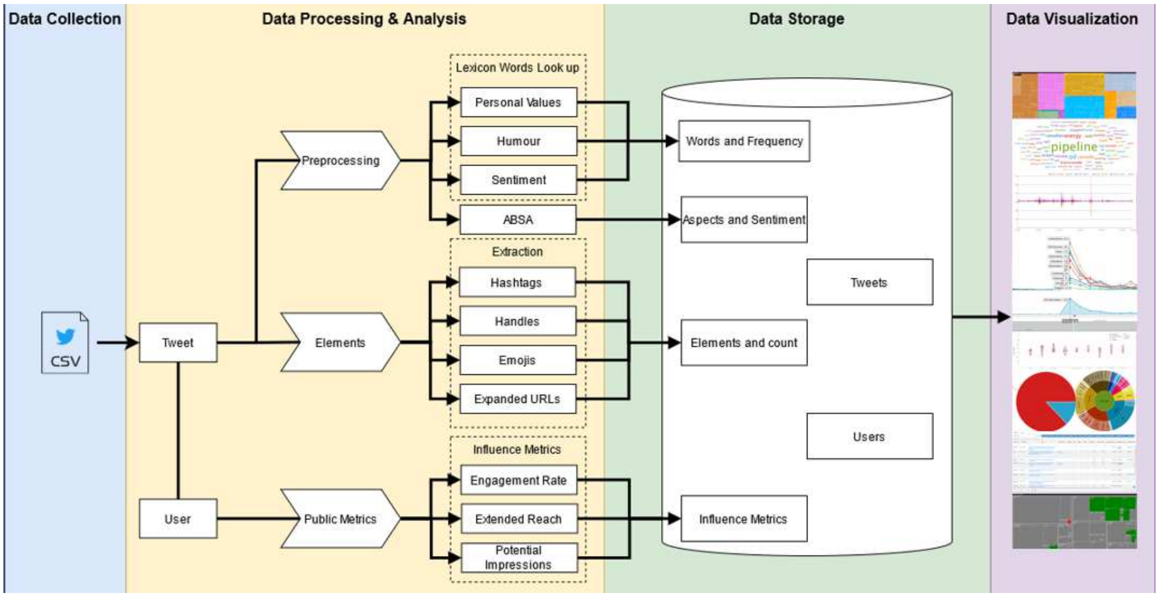
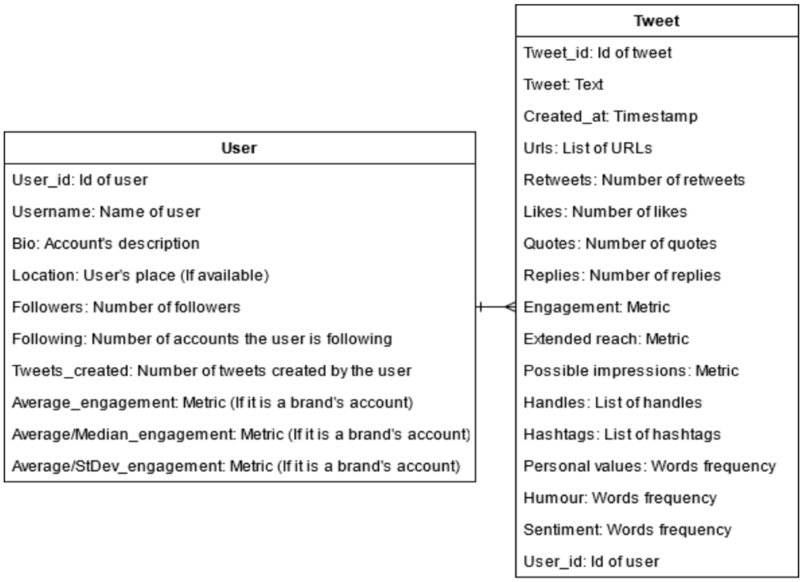


Figure 3.1: The Data Processing Pipelines in *SoDa-TAP* .



User
User_id: Id of user
Username: Name of user
Bio: Account's description
Location: User's place (If available)
Followers: Number of followers
Following: Number of accounts the user is following
Tweets_created: Number of tweets created by the user
Average_engagement: Metric (If it is a brand's account)
Average/Median_engagement: Metric (If it is a brand's account)
Average/StDev_engagement: Metric (If it is a brand's account)

Tweet
Tweet_id: Id of tweet
Tweet: Text
Created_at: Timestamp
Urls: List of URLs
Retweets: Number of retweets
Likes: Number of likes
Quotes: Number of quotes
Replies: Number of replies
Engagement: Metric
Extended reach: Metric
Possible impressions: Metric
Handles: List of handles
Hashtags: List of hashtags
Personal values: Words frequency
Humour: Words frequency
Sentiment: Words frequency
User_id: Id of user

Figure 3.2: Users and Tweets.

post to run the analysis through *SoDa-TAP* . To that end, *SoDa-TAP* supports three different alternative methods:

1. **Custom CSV File.** *SoDa-TAP* is capable to work with custom created datasets as a CSV file. The only condition that this file should comply with is that it requires a header and subsequent rows need to be well formatted.

2. **Official API Client.** *SoDa-TAP* includes a custom client that access the full-archive search API[2] endpoint from Twitter. To use this client, a valid Twitter's Research Bearer Token must be provided, to access historical tweets beyond the default seven days that Twitter provides without a research account.

3. **Scraper.** *SoDa-TAP* incorporates Twint[3], that allows to fetch tweets without the need of an API research key. It is important to note that this method relies heavily on Twitter's front-end website, which means that the limitations of this package are ruled by Twitter's constraints that are put on their website. Hence, the data collected will differ from the one available through the official API.

4. **Tweet ID Hydration.** A typical use case of social-platform data analysis involves to use already created datasets shared online. Twitter allows the curation of such datasets only as collections of tweets' IDs. That is why *SoDa-TAP* incorporates a pipeline developed with the help of twarc2[4], which is used to read a text file composed of multiple IDs written per line.

A transformation function is applied to the retrieved data, depending if it comes from the API client or the tweet hydration process. This is done to parse the results that come as JSON objects and to select the required fields that are later stored in a CSV format. As a final step, after the execution of any of the previously described methods, a CSV file is created, which is later used for further processing and analysis.

3.1.2 Data Processing and Analysis

The core of *SoDa-TAP* resides in this component that includes the functions that process the collected data. This component was developed having in mind flexibility, scalability, reliability and automation, which is why we implemented it using Kafka

[2] https://developer.twitter.com/en/docs/twitter-api/tweets/search/api-reference/get-tweets-search-all

[3] https://github.com/twintproject/twint

[4] https://twarc-project.readthedocs.io/en/latest/twarc2_en_us/

[69] and Spark Structured Streaming [73]. Kafka handles new incoming data, while Spark at the same time, is consuming it and is applying all of the transformations, and metrics.

Text Analysis

Three modules are responsible for text-analysis functionalities. The module responsible for dictionary look up and Aspect-Based Sentiment Analysis (ABSA) first checks if there is any curated or custom dictionary created that needs to be transformed into the expected *SoDa-TAP* format, of value pairs consisting of (a) a word and (b) a number value, if the words are associated to ranges. If the dictionary associates words into classes, the first value defines the word and the second value its category, where the second word value will be lemmatized, for example, if the word was "playing", it will be normalized to "play". For both types of dictionaries, each value pair is expected in a single line text file, that will later be transformed into a JSON file.

Preprocessing

The first step of the overall process involves a text-preprocessing pipeline that removes unnecessary elements from each input post. For the dictionary look up, the tweet-preprocessor[5] and gensim[6] libraries are used to remove stop words, handles, URLs, emojis, numbers, hashtags, punctuation and non-unicode characters and transform all words into lowercase. Next, the lemmatization module from the Natural Language Toolkit (NLTK[7]) package is invoked, to convert all the text into its base form.

Next, each post is processed to extract word frequencies, depending on the already curated dictionaries that are available in the system, in this case: (a) personal values; (b) sentiment; and (c) humour. By the time the system initializes, it loads the set of dictionaries that are declared from the configuration file. Such dictionaries are

[5] https://pypi.org/project/tweet-preprocessor/

[6] https://radimrehurek.com/gensim/

[7] https://www.nltk.org/

frequently consulted each time their respective pipeline is executed. The following are their characteristics.

(a) **Personal Values.** A list of words proposed by Ponizovskiy, Ardag, Grigoryan, Boyd, Dobewall, and Holtz [108], grouped in ten categories that denote personal values: self direction, stimulation, hedonism, achievement, power, security, conformity, tradition, benevolence, and universalism.

(b) **Sentiment.** A list of words proposed by Hu and Liu [109], divided in two groups: positive and negative.

(c) **Humour.** A list of words proposed by Engelthaler and Hills [110], useful for other works [111]. Based on four ranges of intensity: (1) Super Funny, 4 to 5; (2) Medium Funny, 3 to <4; (3) Funny, 2 to <3; and (4) Not Funny, 1 to <2.

Aspect-Based Sentiment Analysis

For the purposes of ABSA, numbers, URLs and handles are removed, leaving punctuation and emoticons/emojis intact, since sentiment analysis tools view them as a contribution to emotion. Later, all text is transformed into lowercase, except those with all letters capitalized, since sentiment analysis tools also view this convention as a contribution to emotion. Hashtags are split into words with CrazyTokenizer [112] and are left as part of the text. The ABSA pipeline starts by applying dependency parsing, where we search for keywords in the text that match with our custom list of energy related words extracted from the Natural Resource Canada's Energy Factbook, the energy industry keywords and project, Wikipedia pages, and related terms associated to the previous words that were found in ConceptNet [113]. When aspects are found, we extract at the same time their span within the text with FlashText's [114]. This span would later be used to locate words around the aspect, to determine if such words alter the aspect's meaning. More in detail, the words around each aspect found in the text are evaluated to see if they carry an emotion that could impact the use of the aspect in the text. When there are multiple aspects in the text, they are separated alongside the words around them for an independent evaluation as mentioned before. When there are no emotive words around the aspect, a window of three words on the left and right from the aspect are selected. Next, VADER [115] is used to extract sentiment, since it is known to work well with social media text. We pick the compound polarity from the result of VADER's sentiment function that is applied to the extracted text where the aspect was found, which is divided into the following ranges (-1.0,-0.5), (-0.5,0.5), (0.5,1) that denote negative, neutral, and positive sentiment respectively.

Element Extraction

A set of functions can extract different posts' elements. So far, we have included the look up of traditional elements that are more likely to be found in most of the social platforms' post(s) such as hashtags, handles, emojis/emoticons and URLs. At the same time, a frequency function is applied over each result in order to quantify each one of them, so that they can be used to support quantitative/qualitative analyses.

Engagement Calculation

Through the application of standard formulas applied in professional platforms [14, 116, 117], engagement rate, extended reach, and potential impressions based on the authors' post metadata are calculated to quantify the impact of each tweet. Engagement rate can be calculated in two ways, based on retweets and likes (i), or retweets, likes, replies and quotes (ii), divided by the number of the followers of the tweet's author multiplied by 100. Extended reach considers the number of retweets divided by the total number of tweets done by the author, multiplied by 100. The author's potential impressions metric takes as input the number of author's followers multiplied by the total number of the author's tweets count.

Image Analysis

We developed a custom console application for image processing. It includes (a) a client that downloads images from an external server, and (b) a client that requests the image processing analysis and receives as a result a JSON document, including all the image properties. *SoDa-TAP* uses this application to computer the following properties:

- **Color Scheme Analysis:** With colorgram[8] we get the number of color degrees seen in an image as RGB values.
- **Object detection:** With Faster R-CNN Inception ResNet V2[9] we select objects with bounding boxes.
- **Image classification:** Inference is done through the model inception v3[10] that returns a confidence value of any element that is present in the image (i.e. bottle, street sign, etc.).
- **Sentiment Analysis:** Sentiment percentage score is done through the model VGG19 finetuned with the Twitter for Sentiment Analysis (T4SA) dataset [118] within 3 degrees (negative, neutral and positive).
- **OCR:** With pytesseract[11] we do recognition of text (if available).
- **Face Analysis:** With deepface [119] we do recognition of faces (if available).

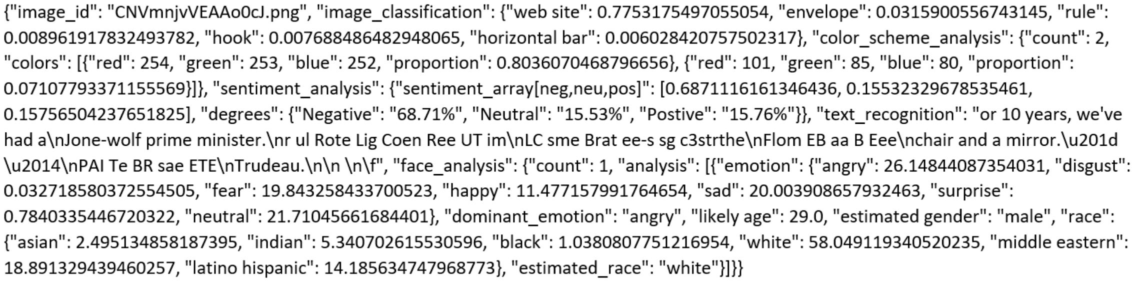
One example of the analysis results can be seen in Figure 3.3. An advantage of this API, is that an available GPU can be adapted, in order to enhance the processing time of images.

[8] https://pypi.org/project/colorgram.py/

[9] https://tfhub.dev/google/faster_rcnn/openimages_v4/inception_resnet_v2/1

[10] https://tfhub.dev/google/imagenet/inception_v3/classification/5

[11] https://pypi.org/project/pytesseract/



```
{"image_id": "CNVmnjvVEAAo0cJ.png", "image_classification": {"web site": 0.7753175497055054, "envelope": 0.0315900556743145, "rule":
0.008961917832493782, "hook": 0.007688486482948065, "horizontal bar": 0.006028420757502317}, "color_scheme_analysis": {"count": 2,
"colors": [{"red": 254, "green": 253, "blue": 252, "proportion": 0.8036070468796656}, {"red": 101, "green": 85, "blue": 80, "proportion":
0.07107793371155569}]}, "sentiment_analysis": {"sentiment_array[neg,neu,pos]": [0.6871116161346436, 0.15532329678535461,
0.15756504237651825], "degrees": {"Negative": "68.71%", "Neutral": "15.53%", "Postive": "15.76%"}}, "text_recognition": "or 10 years, we've
had a\nJone-wolf prime minister.\nr ul Rote Lig Coen Ree UT im\nLC sme Brat ee-s sg c3strthe\nFlom EB aa B Eee\nchair and a mirror.\u201d
\u2014\nPAI Te BR sae ETE\nTrudeau.\n\n \n\f", "face_analysis": {"count": 1, "analysis": [{"emotion": {"angry": 26.14844087354031, "disgust":
0.032718580372554505, "fear": 19.843258433700523, "happy": 11.477157991764654, "sad": 20.003908657932463, "surprise":
0.7840335446720322, "neutral": 21.71045661684401}, "dominant_emotion": "angry", "likely age": 29.0, "estimated gender": "male", "race":
{"asian": 2.495134858187395, "indian": 5.340702615530596, "black": 1.0380807751216954, "white": 58.049119340520235, "middle eastern":
18.891329439460257, "latino hispanic": 14.185634747968773}, "estimated_race": "white"}]}}
```

Figure 3.3: Example of Image Processing API Result.

URL Expander

There are cases that the URL extracted from a post is not fully expanded. In Twitter when we request for the expanded URL, there might be cases that it is shortened by an external service other than Twitter's. For that reason, we incorporated an efficient tool to expand URL(s) called urlExpander [120]. It takes care of launching the desired amount of threads and it creates a caching file to prevent requesting the same URL twice. The results are later written in a JSON file as a single line, for later mapping and processing with their respective post.

3.1.3 Implementation of the Data Processing and Analysis Component

Docker containers are used to make the system easy to deploy and adapt. Kafka, as the data transfer engine, requires two other important containers to help it move the data from one place to the other: (i) zookeeper [121], a service to keep track of Kafka's internal functionalities; and (ii) schema-registry[12], a service to save the structure of the data that Kafka is transporting. These containers need to be running at the same time as Kafka's server container, to be able to design a data transfer pipeline. Spark, on the other hand, is simply downloaded locally into the virtual environment created with Vagrant, as previously introduced. The following are key concepts that are needed to understand the end-to-end data processing.

[12] https://docs.confluent.io/platform/current/schema-registry/index.html

Kafka is defined in terms of the following:

- **Broker:** A Kafka cluster server, responsible of making the connection fault tolerant.
- **Topic:** A message buffer that Kafka uses to transport bytes of data. It has the capability to be deployed and distributed in a Kafka's cluster server.
- **Producer:** An external service that contributes with new data and is capable to send it over a Kafka topic.
- **Consumer:** An external service that can range from reading a topic to storing the data or to applying any other set of operations.
- **Connector:** A package developed to link Kafka with an external service and let them communicate.

Spark requires the definition of the following:

- **Context:** The initialization instance that allows Spark to start with all the required configurations such as: number of cores and RAM to use, libraries to link, etc.
- **Data Partition:** Techniques used to distribute data in different chunks through the available cores. A golden rule is to multiply the number of cores times four, to designate the partition number.
- **DataFrames:** A table to store the data and to apply transformations over it. This data is divided between the processors and depend on the data partition that reign how fast the transformations will be executed.

Defining a Schema for Data Ingestion

The collected data, available as a CSV file, is ready to be sent over to the *Data Processing and Analysis* component as seen in Figure 3.1. But, one of the conditions and standard rule to start processing data is to know its schema. *SoDa-TAP* is able to automatically generate this schema on the fly, through an intermediary step before this data is fully sent over to Spark for processing, and this is achieved thanks to the compatibility between the two. The first element from a Kafka topic is extracted, then Spark reads its value and parses its schema as a JSON. This JSON file is later used to read all the entries available within the topic, in order to start the processing and the analysis.

Kafka and Spark in Action

In *SoDa-TAP* , Kafka and Spark are always alive processes. Our motivation for building *SoDa-TAP* on top of them was their flexibility, capability and compatibility to work with a vast number of other technologies. Both of the tools support the system with three noticeable features:

1. **Multiple Data Sources.** Kafka has a big catalog of plugins to connect as a data source almost any type of data, API or database.
2. **Data Transfer Adapters.** Spark has native connectors that support the connectivity and transfer of data to external systems, but also, if a connector is not available natively, there is a chance that it can be installed as an extra package (in the case of using the python language).
3. **Easy Scale Up and Down.** Spark and Kafka have a set of configuration options that can be adjusted depending on the computing resources that are available.

The starting point of the system begins when Spark reacts to the new incoming data that Kafka is reading. This is achieved through the use of the "Spooldir Source Connector"[13], a Kafka source that is constantly checking for the presence of CSV files

[13] https://www.confluent.io/hub/jcustenborder/kafka-connect-spooldir

that have not been processed yet. A requirement of this connector is to have three directories that are constantly in use: (1) unprocessed, to contain all the files that are about to be processed; (2) error, to contain all the files that were not able to be processed; (3) processed, to contain all the files that were able to be processed. The most important folder is "unprocessed", once this connector detects a new file in this folder, it starts reading it into a previously configured Kafka topic. The reason of using this connector is to give the end user the facility to keep adding more files into the folder. If the case is that a new file is moved into it, it will be instantly processed, without the need to stop or restart the engine.

While the file is being read into the Kafka topic, Spark comes into the picture. The initial step to do processing and analysis starts by reading the schema stored as a JSON of the data to be processed, as previously discussed. This schema is used by Spark to read the data that is coming from the Kafka topic. Around 100,000 records (a number that can be scaled) are read from the topic until it reaches the end of the file. Since each record comes in the form of bytes, Spark needs to do the conversion from bytes to string, and from string to JSON. Next, the incoming data starts to be incrementally appended into a Spark DataFrame, to later be able to apply the set of custom functions designed for each of the pipelines described earlier. To finalize, *SoDa-TAP* takes advantage of Pandas[14] and SQLAlchemy[15] to enable a communication with the database and to automatically generate data schemas from the processed data. For this, each 10,000 new records (a number that can be adapted to fit another use case) inserted in the DataFrame, are sent over to the *Data Storage* component.

3.1.4 Data Storage

Low latency, high availability and a powerful query interface are what we wanted to deliver alongside a scalable processing engine. This type of data has the potential

[14] https://pandas.pydata.org/

[15] https://www.sqlalchemy.org/

to increase exponentially in volume, and have many keys and values associated to it. Therefore, we incorporated CrateDB [80] as the designated database for the *SoDa-TAP* data, because, when compared against other open-source databases, this one excelled on all the features and internal characteristics that it offers, including the following.

1. **Distributed and Sharded.** It is able to distribute the data volume over multiple instances of the database and the indexing of the tables are distributed in an optimized way.

2. **Columnar.** Tables are efficiently accessed, since it works over the selection of the necessary columns. Also, they can hold as many columns as needed.

3. **SQL Interface.** SQL language is the default query interface to do transactions in the database and the majority of the traditional operations are available for querying the data. PostgreSQL wire Protocol[16] is present in it and that allows it to be compatible with most of the external applications that can work with a PostgreSQL[17] database.

SoDa-TAP includes two Docker containers for its storage component: CrateDB and FastAPI, to support scalable and efficient data storage and exposure. The following are key considerations to be able to deploy them at scale.

The **CrateDB** configuration involves the following decisions:

- **Nodes.** Fault tolerance is achieved when deploying a minimum of three services of the database.

- **Heap Size.** To be able to select the required data, memory needs to be allocated and it depends on the query that is being performed. A minimum of two gigabytes is recommended per node, to handle this type of tasks.

[16] https://www.postgresql.org/docs/current/protocol.html
[17] https://www.postgresql.org/

- **Ports.** The exposure of only the necessary ports from the main node are needed to limit the access.

The **FastAPI** configuration requires the specification of the following:

- **Port.** The selection and exposure of a non traditional port is desired to give access to other services into the data.

- **Pagination.** To be able to control memory allocation for data transfer, a pagination methodology is necessary to get progressive data loads, if the case is that it is high in size.

CrateDB, FastAPI, Socket.IO and SQLAlchemy in Action

After processed through Spark, data is transported to the CrateDB database, where the "tweets" table is created to hold all incoming data. This table is is responsible to hold all the attributes as denoted in Figure 3.7a. A very useful characteristic of CrateDB is that it allows to create views (they have the same functionality just like tables, but the difference is that they are a projection of a selection of columns of a table). Through the view functionality we are able to execute a SQL query to extract distinct users seen in the original tweets table as shown in Listing 3.1. After the execution of the previous query, a new view named users is created and ready to be queried as a normal table through the exposed API services.

Listing 3.1: SQL Query to Create View of Authors.

```
CREATE VIEW {authors} AS SELECT DISTINCT author_id ,
    author_username , author_bio , author_followers_count ,
    author_following_count , author_tweet_count , max(created_at
    ) FROM {table} GROUP BY author_id , author_username ,
    author_bio , author_followers_count , author_following_count
    , author_tweet_count;
```

SoDa-TAP exposes all the data available in the database through two API services that rely on SQLAlchemy as the middleware for processing SQL transactions, as seen in Table 3.1. One that was developed with FastAPI[18], to support specific data access

[18] https://fastapi.tiangolo.com/

granularity and functionality, and another one that was developed with Socket.IO[19], event-driven oriented, to request progressive batches of data that are triggered by SQL statements through a key set pagination. These implementations were developed with two use cases in mind: (i) an external application only needs to select a couple of rows from a table, accessed through normal HTTP requests; and (ii) an external application needs to consume all the rows of a table, through the use of Socket.IO's library import and its custom functions' definition for the web environment.

Table 3.1: Differences between FastAPI and Socket.IO APIs.

API	**Description**
FastAPI	A set of endpoints for a specific task and control of data access: (a) select data by pieces; (b) select data based on an offset; (c) select data by id(s); and (c) select time ranges of data. The calls need to be done by block and waiting mechanisms to control data transfers.
Socket.IO	A main event-driven function to handle external SQL queries to select data. Each request listens to a query instruction, accompanied with an auto-generated ID. The result is transferred based on the assigned ID to differentiate the client requesting for data, which makes it reliable to continuously listen to frequent calls.

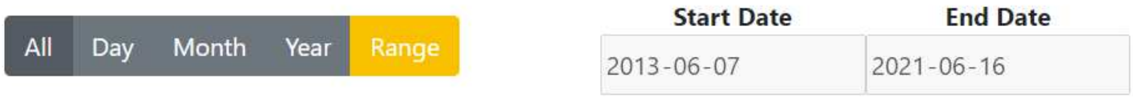
3.1.5 Data Visualization

The current catalog of visualizations available in *SoDa-TAP* , built and configured with ZingChart [91], is divided in two groups. The Dictionary-Based Analysis group was designed and developed to count the frequency of words expressed in text, obtained from applying the lexicon look up pipeline. The ABSA group was designed and developed to count, based on a time frame, the frequency of aspects and terms found in text, to visualize their sentiment association. A special filtering bar as the one shown in Figure 3.4 was developed with Bootstrap[20], to apply a filter to: (a) ag-

[19] https://socket.io/

[20] https://getbootstrap.com/

gregate all data; (b) compare per day, month or year dates or (c) custom dates range. This filtering option is limited to some of the charts, since it was only integrated to those ones that would benefit from showing more granularity. In this section, we cover details about each visualization presented in the Tables 3.5 and 3.6.



Start Date: 2013-06-07

End Date: 2021-06-16

Figure 3.4: Filtering Options.

Dictionary-Based Analysis Visualizations

- **Treemap [Figure 3.5a]:** This chart was developed to visualize words frequency associated to personal values. A big box denotes a personal value and the internal ones are the words that matched the group of the personal value. Each big box has been assigned a specific color that the internal ones inherit. Each box type holds a text label, whether is a personal value group or is the word that was found. Internal boxes can be of different sizes, this attribute defines the total count of words found. This chart can use the filtering option, its use changes the color of the internal boxes into red or green. Green if the case is that the word increased in frequency by comparing the start against the end date, and red if it's the opposite. If the case is that there is no change in the frequency, the box will keep the starting color. Hovering over the boxes to see their volume count is enabled.

- **Line and Area [Figure 3.5b]:** This chart shows cumulative word count per day, depending on the dictionary selected, in order to show data distribution over time, which can lead to discover important dates. Navigation controls such as drag, zoom in or out and hovering over data points are enabled.

- **Streamgraph [Figure 3.5c]:** This chart has similar functionality as the line and area chart. Its difference lays in the visual distribution of the data, which can lead to discover word volume in a faster way. Navigation controls such as drag, zoom in or out and hovering over data points are enabled.

- **Words, Emojis, Hashtags and Handles Cloud [Figures 3.5d, 3.5e, 3.5f, 3.5g]:** This chart allows to see in a graphical way the words, emojis, hashtags or handles that have the highest frequency count. Hovering over the elements to see their volume count is enabled.

Aspect-Based Sentiment Analysis Visualizations

- **Sunburst [Figure 3.6a]:** This chart aggregates terms to their respective aspects, this can be noticed as a slice that represents their frequency count. This chart can use the filtering option and by hovering over the elements, their volume count is displayed.

- **Pie [Figure 3.6b]:** This chart mentions the amount of tweets that have an aspect and the amount that do not. Hovering over the slices to see their volume count is enabled.

- **Violin [Figure 3.6c]:** This chart shows the distribution of each aspect mean sentiment compound. By hovering over individual mean compound values per violin, it shows its value.

- **Treemap [Figure 3.6d]:** This chart was developed to visualize words frequency associated to aspect terms and their sentiment. A big box denotes an aspect and the internal ones are the terms associated with the aspect. Each box has a color that depends on the sentiment of the aspect: (a) red, negative; (b) grey, neutral; and (c) green, positive. Each box type holds a text label, whether is an aspect or is a term. Internal boxes can be of different sizes, this attribute defines the total count of terms found. This chart can use the filtering option and by hovering over the boxes their volume count is shown.

Data Table Exploration

Beyond the above visualizations, *SoDa-TAP* supports a simple and intuitive tabular data-exploration component developed with FancyGrid[21]. We currently use it to show two tables, tweets and users. Both tables hold same functionalities, but their main difference are the columns associated with their data and exclusively, the table of tweets aggregates: the average, mean and standard deviation of engagement. The following are the key functionalities of them, and Figure 3.7c shows their appearance:

- **Table Search.** Filter the table based on a keyword, sentence or pair of words throughout the table. Only the rows that match with the search criteria will be visible.

- **Column Selection.** Show or hide specific columns to only pay attention to the ones that are required.

- **Column Filtering.** Filter based on a keyword in case of a text column. If the case is that is a numeric column, different operators: $<, >, \leq, \geq, \neq$; can be used to filter based on a comparison of the desired value against the whole column.

- **Data Export.** Export all the table's data into a CSV file and download it.

In general, links can be clicked and a navigation menu is present to: select the number of rows to show (number that is configurable through a drop-down menu), move between pages, see the current page number and the total amount of rows available.

3.2 Statistical Analysis Toolkit

SoDa-TAP includes a component that facilitates and automates the statistical analysis of the correlations of different data features with the engagement metrics. To

[21] https://fancygrid.com/

make this possible, a Jupyter notebook was designed with all the necessary tools and scripts to configure it based on custom use cases. The following are all the available features:

- **Fast Analyses.** Modin[22] with Ray[23], a parallel DataFrame to speed up Pandas is used as a processing engine to do all calculations and analyses.

- **Visualizations.** Currently, box plots are used as the main form to visualize elements and engagement groups relationship.

- **Binning.** A function was developed to automate the creation of groups that can be adapted to custom ranges.

- **Custom Variables Definition.** A custom defined JSON like dictionary as seen in Figure 3.8, was developed to personalize it with chosen: dependent and independent variables, binning names and grouping labels, to later use it for custom analyses.

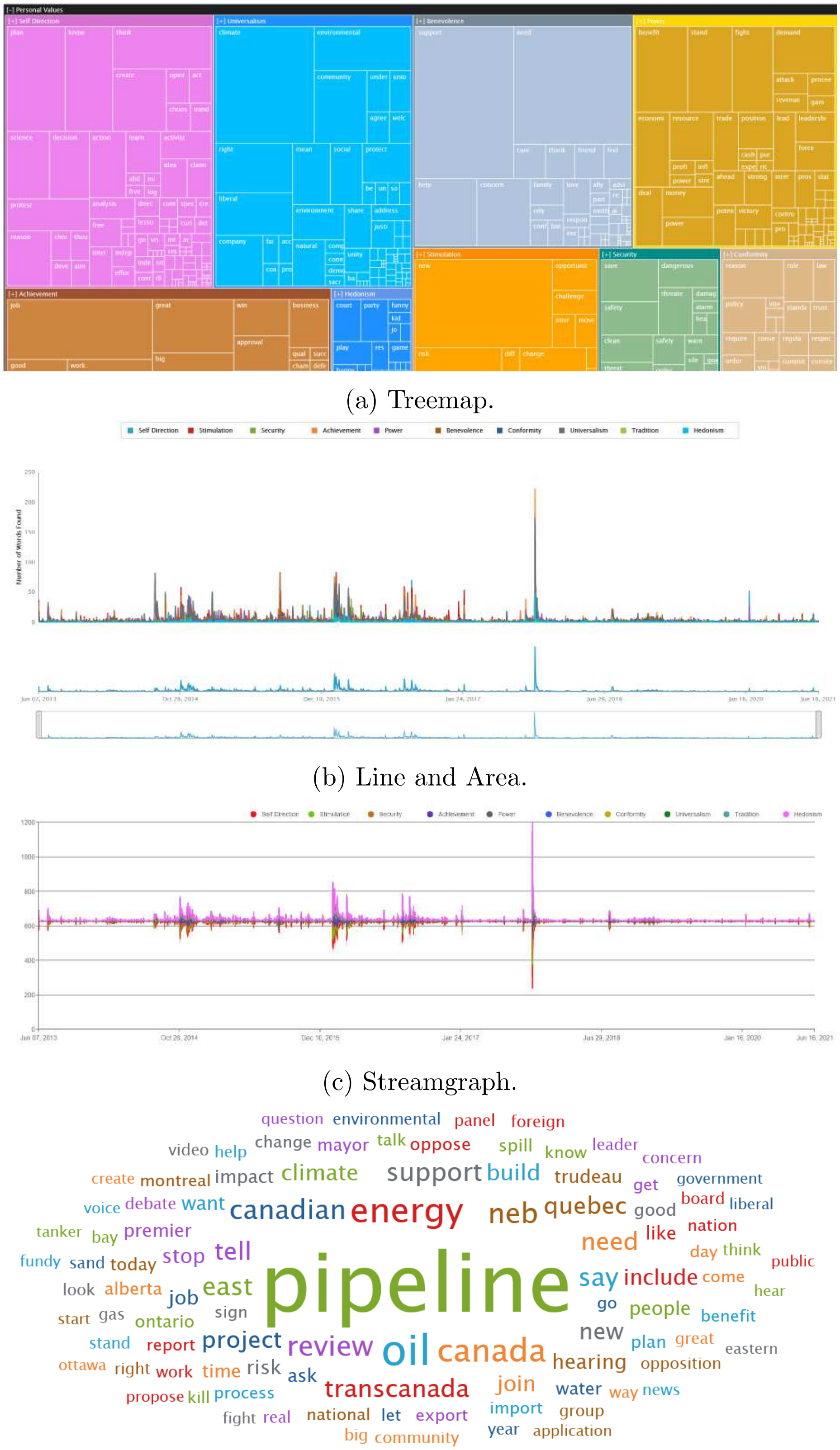
- **One and N-Way ANOVA.** The calculation of one and n-way ANOVA was developed to discover the relationship of dependent and independent variables. This is possible thanks to the package named statsmodels[24] and matplotlib[25].

[22] https://modin.readthedocs.io/en/stable

[23] https://www.ray.io/

[24] https://www.statsmodels.org/stable/index.html

[25] https://matplotlib.org/



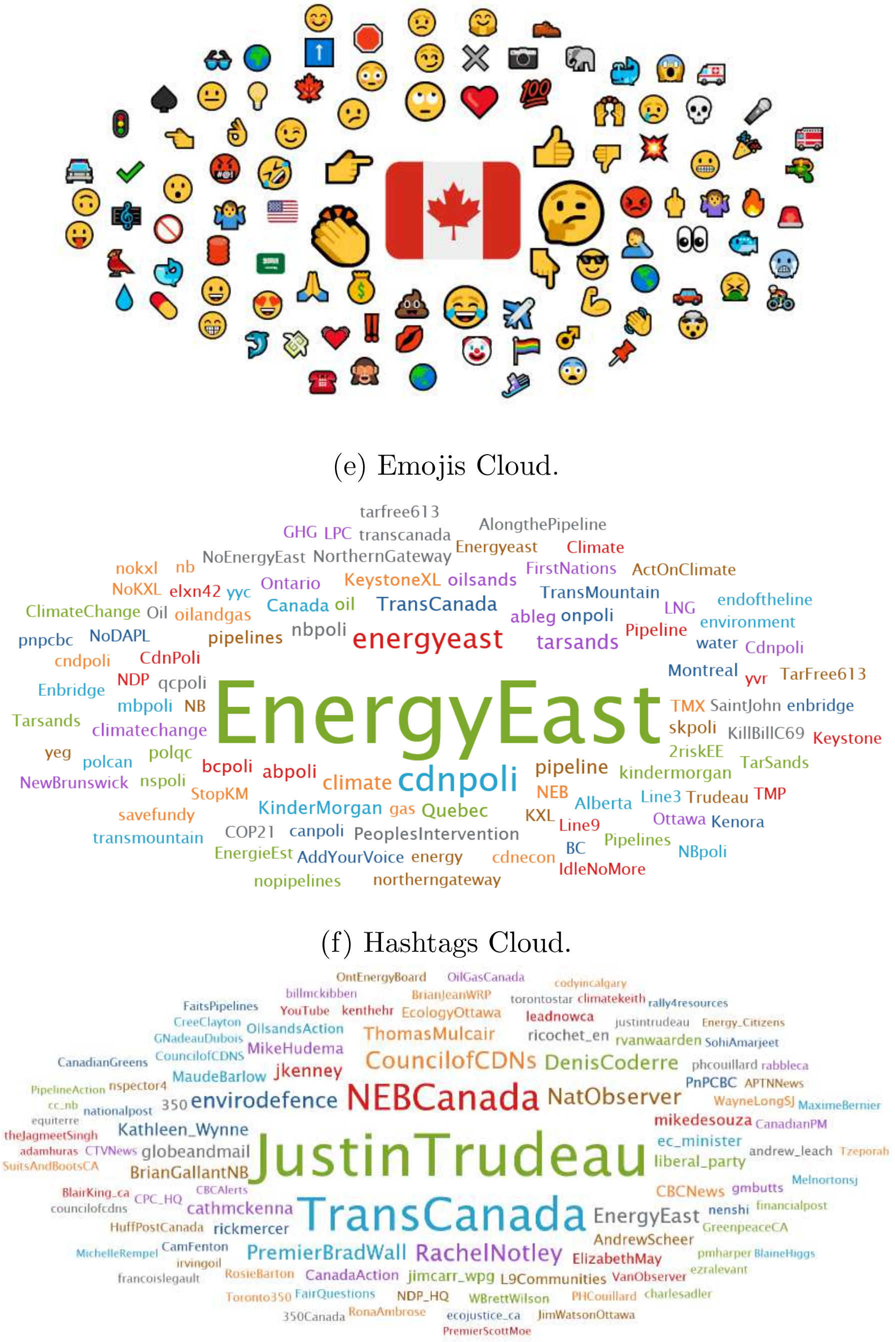
(a) Treemap.

(b) Line and Area.

(c) Streamgraph.

(d) Words Cloud.

Figure 3.5: Dictionary-Based Analysis Charts.

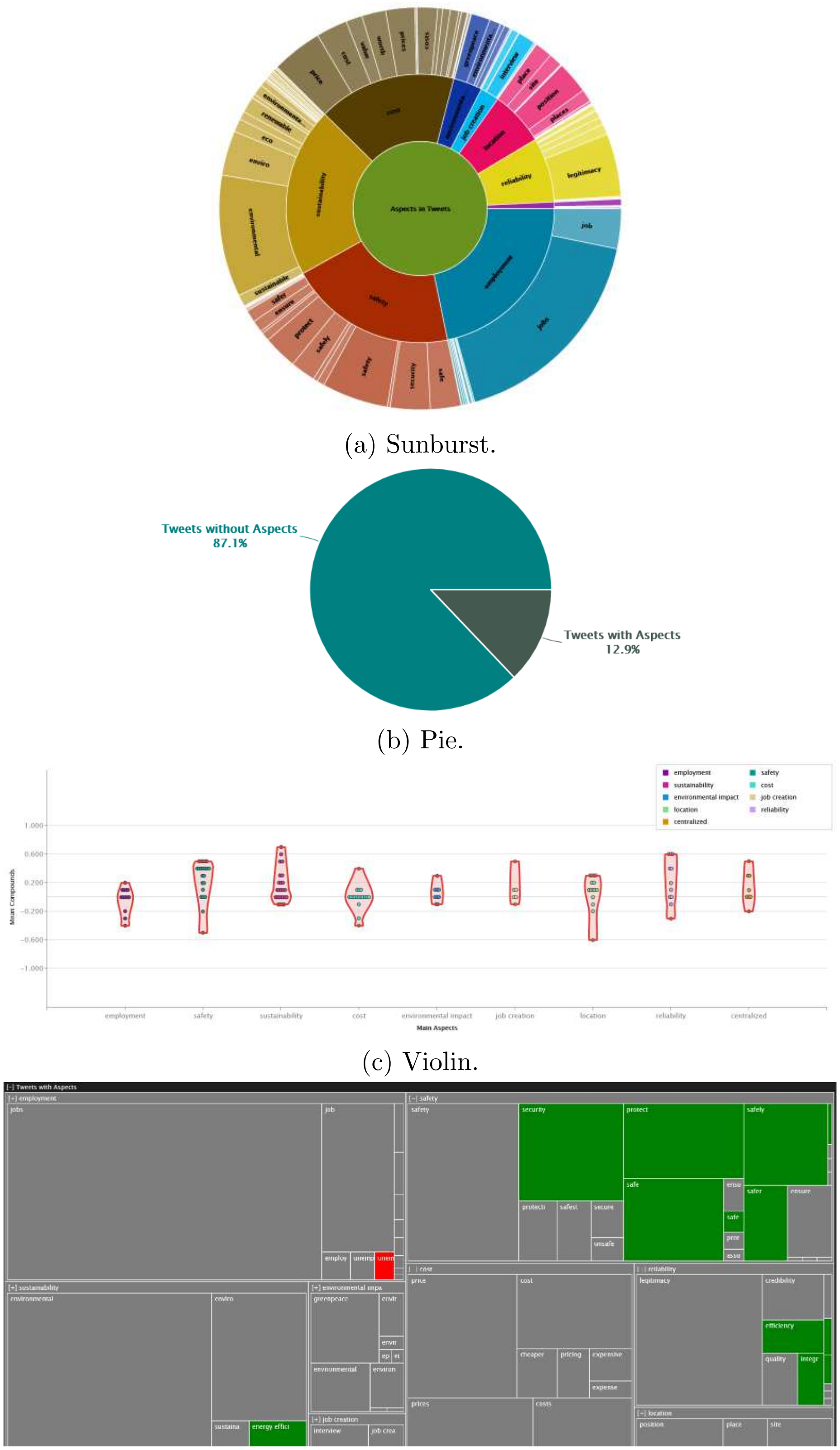


(e) Emojis Cloud.

(f) Hashtags Cloud.

(g) Handles Cloud.

Figure 3.5: Dictionary-Based Analysis Charts (cont.).



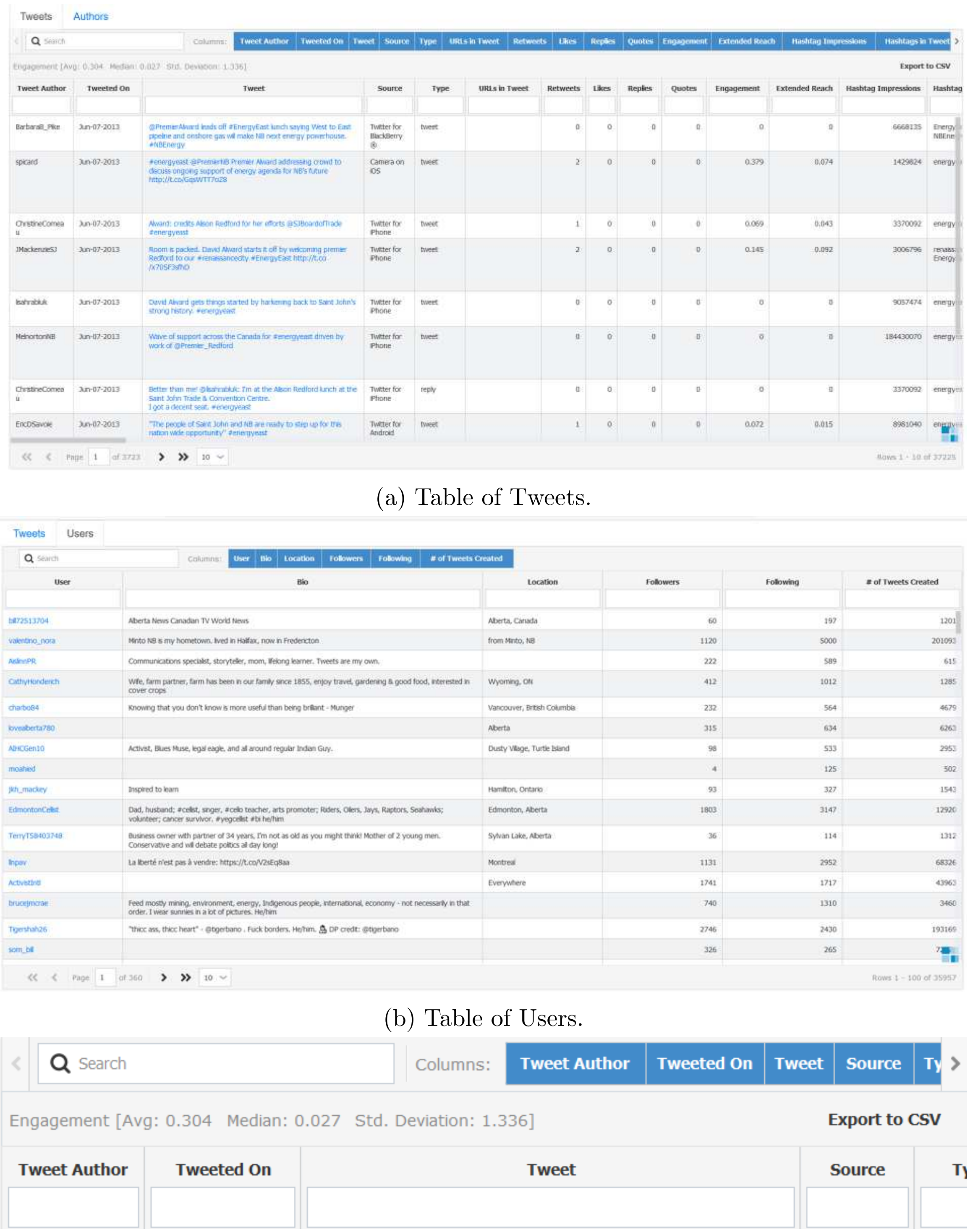
(a) Sunburst.

(b) Pie.

(c) Violin.

(d) Treemap.

Figure 3.6: Aspect-Based Sentiment Analysis Charts.



Engagement [Avg: 0.304 Median: 0.027 Std. Deviation: 1.336]

Tweet Author	Tweeted On	Tweet	Source	Type	URLs in Tweet	Retweets	Likes	Replies	Quotes	Engagement	Extended Reach	Hashtag Impressions	Hashtag
BarbaraB_Pike	Jun-07-2013	@PremierAlward leads off #EnergyEast lunch saying West to East pipeline and onshore gas will make NB next energy powerhouse. #NBEnergy	Twitter for BlackBerry ®	tweet		0	0	0	0	0	0	6668135	Energy NBEne
spicard	Jun-07-2013	#energyeast @PremierNB Premier Alward addressing crowd to discuss ongoing support of energy agenda for NB's future http://t.co/GqsWTT7oZ8	Camera on iOS	tweet		2	0	0	0	0.379	0.074	1429824	energy
ChristineComeau	Jun-07-2013	Alward: credits Alison Redford for her efforts @SJBoardofTrade #energyeast	Twitter for iPhone	tweet		1	0	0	0	0.069	0.043	3370092	energy
JMackenzieSJ	Jun-07-2013	Room is packed. David Alward starts it off by welcoming premier Redford to our #renaissancecity #EnergyEast http://t.co/X70SF3sfhD	Twitter for iPhone	tweet		2	0	0	0	0.145	0.092	3006796	renais Energy
lsahrabluk	Jun-07-2013	David Alward gets things started by harkening back to Saint John's strong history. #energyeast	Twitter for iPhone	tweet		0	0	0	0	0	0	9057474	energy
MelnortonNB	Jun-07-2013	Wave of support across the Canada for #energyeast driven by work of @Premier_Redford	Twitter for iPhone	tweet		0	0	0	0	0	0	184430070	energy
ChristineComeau	Jun-07-2013	Better than me! @lsahrabluk: I'm at the Alison Redford lunch at the Saint John Trade & Convention Centre. I got a decent seat. #energyeast	Twitter for iPhone	reply		0	0	0	0	0	0	3370092	energye
EricDSavoie	Jun-07-2013	"The people of Saint John and NB are ready to step up for this nation wide opportunity" #energyeast	Twitter for Android	tweet		1	0	0	0	0.072	0.015	8981040	energye

Page 1 of 3723 10

Rows 1 - 10 of 37225

(a) Table of Tweets.

User	Bio	Location	Followers	Following	# of Tweets Created
bill72513704	Alberta News Canadian TV World News	Alberta, Canada	60	197	1201
valentino_nora	Minto NB is my hometown. lived in Halifax, now in Fredericton	from Minto, NB	1120	5000	201093
AislinnPR	Communications specialist, storyteller, mom, lifelong learner. Tweets are my own.		222	589	615
CathyHonderich	Wife, farm partner, farm has been in our family since 1855, enjoy travel, gardening & good food, interested in cover crops	Wyoming, ON	412	1012	1285
charbo84	Knowing that you don't know is more useful than being brilliant - Munger	Vancouver, British Columbia	232	564	4679
lovealberta780		Alberta	315	634	6263
AIHCGen10	Activist, Blues Muse, legal eagle, and all around regular Indian Guy.	Dusty Village, Turtle Island	98	533	2953
moahied			4	125	502
jkh_mackey	Inspired to learn	Hamilton, Ontario	93	327	1543
EdmontonCellist	Dad, husband; #cellist, singer, #cello teacher, arts promoter; Riders, Oilers, Jays, Raptors, Seahawks; volunteer; cancer survivor. #yegcellist #bi he/him	Edmonton, Alberta	1803	3147	12920
TerryTS8403748	Business owner with partner of 34 years, I'm not as old as you might think! Mother of 2 young men. Conservative and will debate politics all day long!	Sylvan Lake, Alberta	36	114	1312
lnpav	La liberté n'est pas à vendre: https://t.co/V2sEq8aa	Montreal	1131	2952	68326
ActivistIntl		Everywhere	1741	1717	43963
brucejmcrae	Feed mostly mining, environment, energy, Indigenous people, international, economy - not necessarily in that order. I wear sunnies in a lot of pictures. He/him		740	1310	3460
Tigershah26	"thicc ass, thicc heart" - @tigerbano . Fuck borders. He/him. DP credit: @tigerbano		2746	2430	193169
som_bil			326	265	7

Page 1 of 360 10

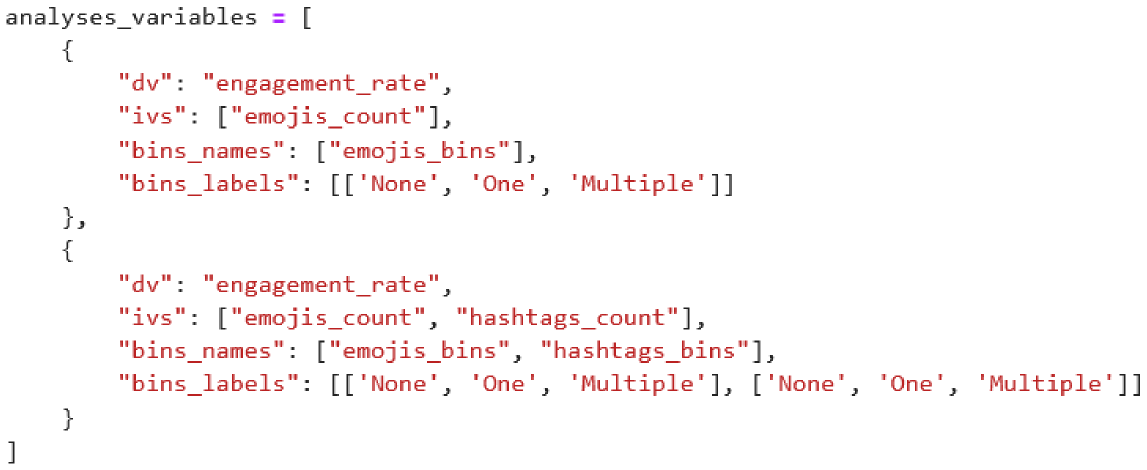
Rows 1 - 100 of 35957

(b) Table of Users.

Engagement [Avg: 0.304 Median: 0.027 Std. Deviation: 1.336]

(c) Filters Available for Table of Tweets.

Figure 3.7: Data Tables.



```
analyses_variables = [
    {
        "dv": "engagement_rate",
        "ivs": ["emojis_count"],
        "bins_names": ["emojis_bins"],
        "bins_labels": [['None', 'One', 'Multiple']]
    },
    {
        "dv": "engagement_rate",
        "ivs": ["emojis_count", "hashtags_count"],
        "bins_names": ["emojis_bins", "hashtags_bins"],
        "bins_labels": [['None', 'One', 'Multiple'], ['None', 'One', 'Multiple']]
    }
]
```

Figure 3.8: Example of Custom JSON like dictionary for ANOVA.

Chapter 4

Deployment of *SoDa-TAP*

Because there are so many use cases for social-media analytics and because in many cases the interested stakeholders do not necessarily have software-engineering or IT background, one of the objectives of this work has been to make the deployment and use of *SoDa-TAP* simple. This is why we developed an automatic-deployment process, guided by a configuration file that defines the data sources, the dictionaries included in the deployment, and the pipelines that are required for the custom analyses.

4.1 Automatic Deployment

The automatic-deployment process, diagrammatically depicted in Figure 4.1, enables the deployment of *SoDa-TAP* on a local machine or in the cloud. We decided to use concepts from the "Infrastructure as a Service" model, inspired by work proposed from Huan [122] that is used in Amazon Cloud for configurable virtual appliances, in order to give the system the flexibility to adapt and simplify its configuration, based on the manual input of computing resources that are available or that are desired to be used. To that end, we make use of three tools to define the internal resources of *SoDa-TAP* .

Vagrant[1]. A virtualization solution that, together with VirtualBox[2], is used as a wrapper that triggers the whole deployment through seven main variables,

[1] https://www.vagrantup.com/

[2] https://www.virtualbox.org/

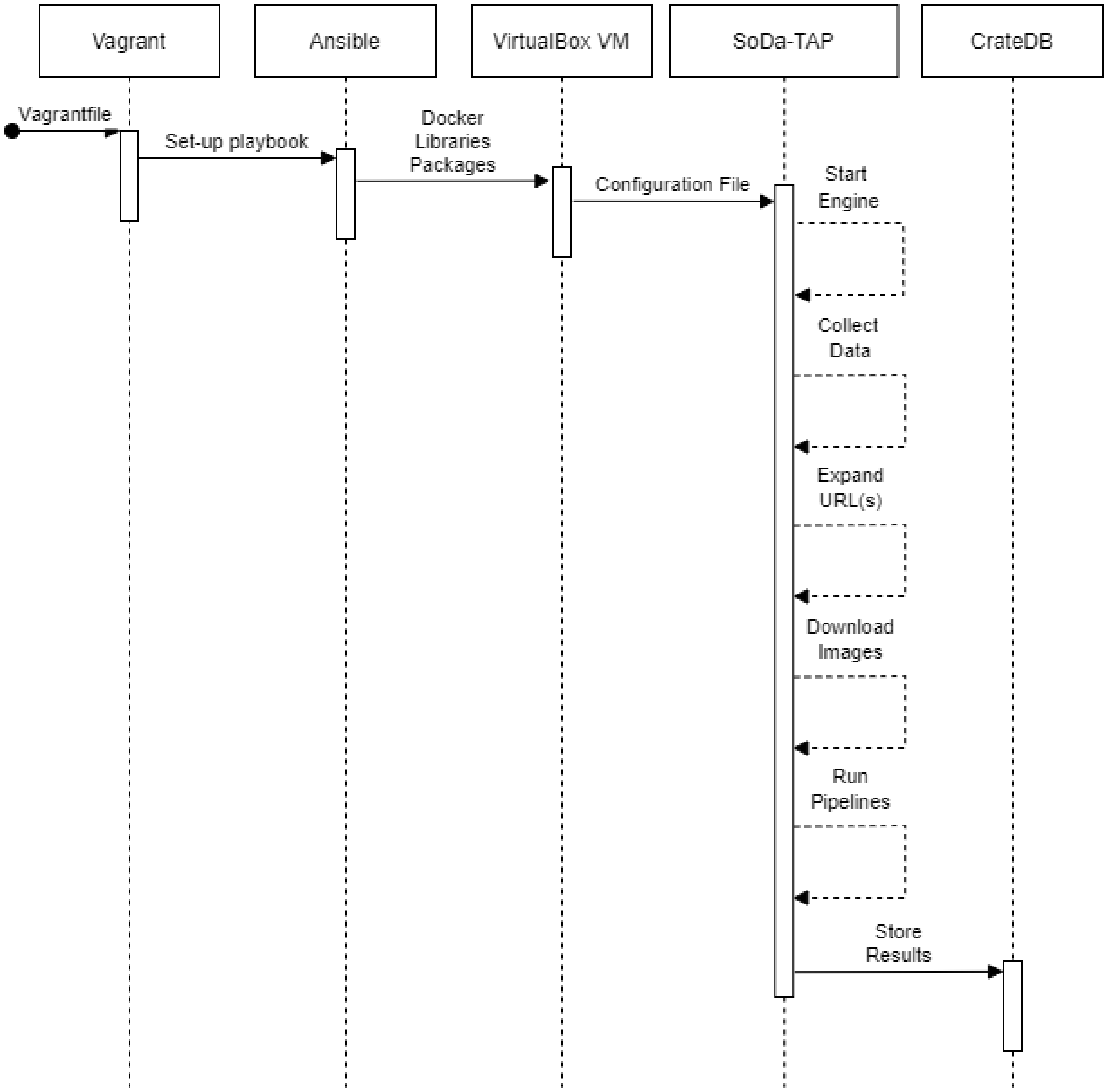


Figure 4.1: Sequence Diagram of *SoDa-TAP* .

as seen in Listing 4.1: (1) “IMAGE_NAME”, an operating system; (2) “N”, the number of virtual machines to launch alongside the master, to create a cluster deployment; (3) “v.memory”, the size of RAM to designate for each virtual machine; (4) “v.cpus”, the number of CPUs that each virtual machine can use; (5) “config.vm.network”, a set of network ports to open for internal/external communication; (6) “master.vm.network”, a custom-defined IP to assign for the virtual machine; and (7) “ansible.playbook”, a file path to a file that will be used to install the required packages and libraries for the system.

Listing 4.1: Vagrantfile.

```
IMAGE_NAME = ”bento/ubuntu−20.04”
#N = 1

Vagrant.configure(”2”) do |config|
    config.vm.provider ”virtualbox” do |v|
        v.memory = {GB}
        v.cpus = {Cores}
    end

    config.ssh.insert_key = true
    config.ssh.forward_agent = true

    # ... potentially many of ...
    config.vm.network ”forwarded_port”,
    guest: {port},
    host:  {port},
    auto_correct: true

    config.vm.define ”SoDa–TAP–vm” do |master|
        master.vm.box = IMAGE_NAME
        master.vm.network ”private_network”, ip:
           ”192.168.50.10”
        master.vm.hostname = ”SoDa–TAP–vm”
        master.vm.provision ”ansible_local” do |ansible|
            ansible.playbook = ”infrastructure−config/
               setup−playbook.yml”
        end
    end
end
```

Ansible[3]**.** An automatic configuration solution that is used alongside Vagrant to help set up all the required libraries and packages that will be needed for *SoDa-TAP* . This is done through a sequence of steps executed once the creation of the virtual machine(s) is finished. An example can be seen in Figure 4.2 and its final outcome is to start all the containers, services, and configurations for *SoDa-TAP* 's components.

Docker[4]**.** This is a containerization solution to package individual components of *SoDa-TAP* . All logic, services and technologies that are mentioned in the *Implemen-*

[3] https://www.ansible.com/

[4] https://www.docker.com/

tation section are defined in a docker-compose YAML file, which are stated as the final execution step from Ansible. The final result from this step, is to start the logic defined from our custom configuration file that *SoDa-TAP* needs to be followed to start a data workflow.

The automatic deployment capability of *SoDa-TAP* can help reduce errors and complexity in the internal behaviour of the system, but most importantly, it can simplify development for the inclusion of other SMA techniques, data sources, data visualizations, and third-party applications consumption. This feature of *SoDa-TAP* and the right selection of tools to achieve it, allows the system to be elastic to a dynamic set of computing resources.

4.2 The Configuration File

A custom defined YAML file was developed to achieve the flexibility needed that gives to the end user the required granular control over the data source and analysis pipelines, so that *SoDa-TAP* is able to execute only the defined and necessary tasks declared by the user. This file as seen and described in Table 4.1, is divided in four sections, which defines the sequential workflow.

1. **Environment:** Is the definition of resources the user would like to provide and that should be available free for *SoDa-TAP* to take over.

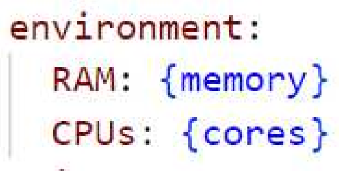
2. **Client:** Is the definition of the data source, which can be a file or a client script to fetch tweets.

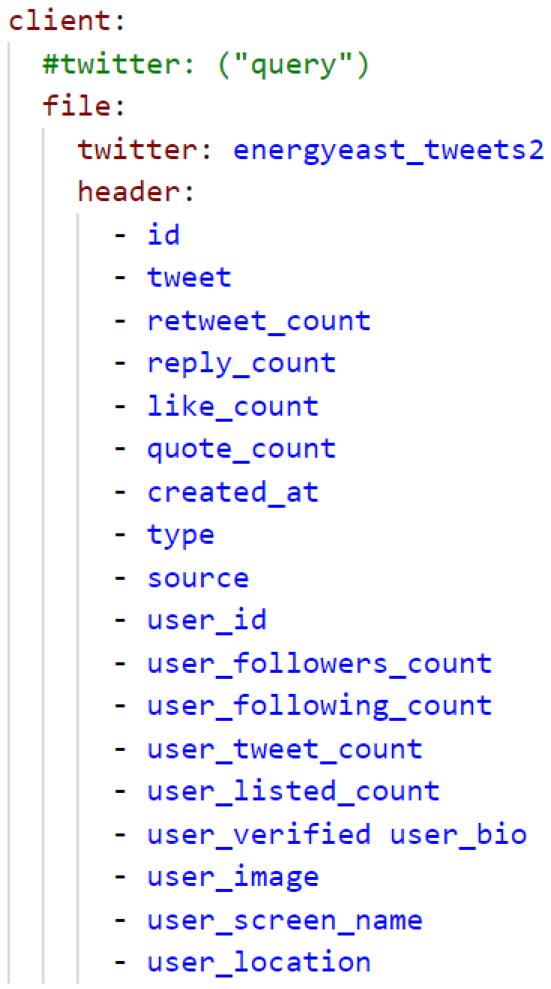
3. **Dictionaries:** Is the definition of the dictionaries that the user would like the text pipelines to execute.

4. **Pipelines:** Is the definition of all the operations and analyses to perform over the data.

Table 4.1: Sections of the Configuration File.

Section	**Description**

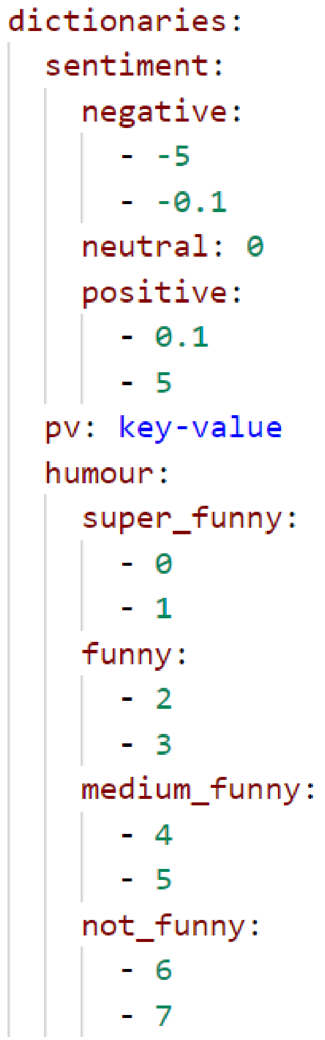

`environment:`
`  RAM: {memory}`
`  CPUs: {cores}`
	• **RAM:** Amount of memory in GB.
	• **CPUs:** Number of processing cores.


`client:`
`  #twitter: ("query")`
`  file:`
`    twitter: energyeast_tweets2`
`    header:`
`      - id`
`      - tweet`
`      - retweet_count`
`      - reply_count`
`      - like_count`
`      - quote_count`
`      - created_at`
`      - type`
`      - source`
`      - user_id`
`      - user_followers_count`
`      - user_following_count`
`      - user_tweet_count`
`      - user_listed_count`
`      - user_verified user_bio`
`      - user_image`
`      - user_screen_name`
`      - user_location`
	• **twitter (or any other social platform name):** A query ad hoc to the social platform to be used to fetch data.
	• **file:** The definition of the filename or path of a CSV file and its header.

Continued on next page

Table 4.1 - *Continued from previous page.*

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>



```
dictionaries:
  sentiment:
    negative:
      - -5
      - -0.1
    neutral: 0
    positive:
      - 0.1
      - 5
  pv: key-value
  humour:
    super_funny:
      - 0
      - 1
    funny:
      - 2
      - 3
    medium_funny:
      - 4
      - 5
    not_funny:
      - 6
      - 7
```

</td><td>

Currently, three dictionaries are embedded to analyze text.

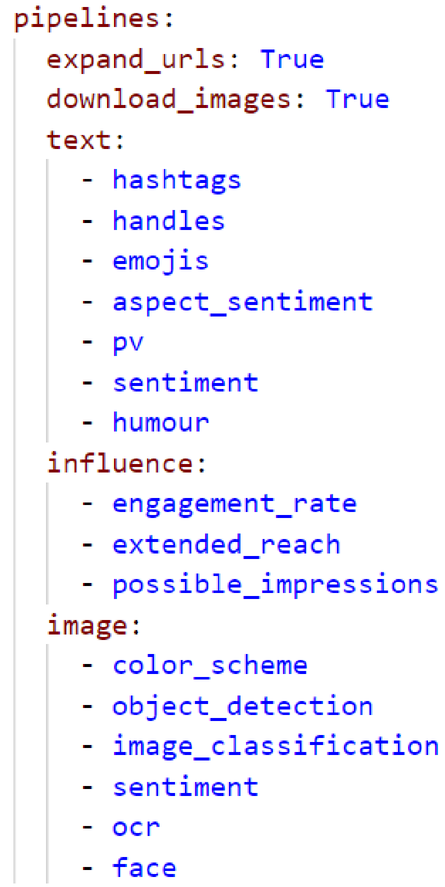
- **sentiment and humour:** Keys as degrees and a number that denotes a range value.
- **pv (personal values):** A word as a key and another word as value that is part of a group.

</td></tr>
</table>

Continued on next page

Table 4.1 - *Continued from previous page.*

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>



```
pipelines:
  expand_urls: True
  download_images: True
  text:
    - hashtags
    - handles
    - emojis
    - aspect_sentiment
    - pv
    - sentiment
    - humour
  influence:
    - engagement_rate
    - extended_reach
    - possible_impressions
  image:
    - color_scheme
    - object_detection
    - image_classification
    - sentiment
    - ocr
    - face
```

</td><td>

- **expand_urls:** Boolean to start a console that is in charge to expand urls in multi-threading.
- **download_images:** Boolean to start a console that is in charge to download images in multi-threading.
- **text:** Set of operations that will be performed over text.
- **influence:** Set of calculations that will be performed over user metadata.
- **image:** Set of analyses that will be performed over images.

</td></tr>
</table>

From the YAML file's available set of variables, a script reads the definitions from "environment" to "dictionaries" and sequentially takes care of the data source and conversion of the dictionaries. "Pipelines" is handled to match the internal variables used within *SoDa-TAP* 's logic. This interaction, makes the further expansion easy, by being able to reuse variables names and components currently available.

4.3 Capabilities of *SoDa-TAP*

A system like *SoDa-TAP* , needs to fulfill the requirements defined by two different scenarios: (i) when there is a boundary of hardware resources, their usage needs to be balanced; and (ii) when there are variations in the incoming volumes of data, the processing workflow should get adapted accordingly. By the leverage of open source tools that are being maintained by official organizations and companies from the software industry, we can be sure that these requirements will be correctly handled, because of the high granularity of configuration options available, that each tool offers to work for a variety of use cases. The following are six capabilities that the selected tools enable for *SoDa-TAP* :

1. **Modular and Extensible:** The use of Docker, to wrap services as containers, makes it possible for *SoDa-TAP* to easily extend functionalities without affecting other modules. At the same time, new services can be paired by taking advantage of network protocols that allows a component to communicate internally or externally.

2. **Integrate Multiple Sources:** Kafka makes it possible for the system to read data that can come from a variety of sources: offline, streaming, API, databases, etc. It has many source connectors available that make it easy to transport data and to build end-to-end data pipelines.

3. **Robust and Scalable:** The integration of Vagrant, Ansible, and Docker allows for the environment of the platform to be flexible, while Kafka, Spark, and CrateDB make the data pipeline fault tolerant and adaptable to data volume.

4. **Analyze Text and Media:** The internal modularity of the logical components of the data processing component helps in the development of tasks that require different resources for data processing according to its data type.

5. **Multiprocessing and Multi-threaded Tasks Support:** The internal modularity of the logical components of the data processing component allows the system to incorporate tasks that not only get benefited from multiprocessing but also from external consumption.

6. **Easy Visualizations Development:** The use of ZingChart and FancyGrid reduces the development time to create highly configurable charts and functionalities.

4.3.1 Expanding on Scalability and Robustness

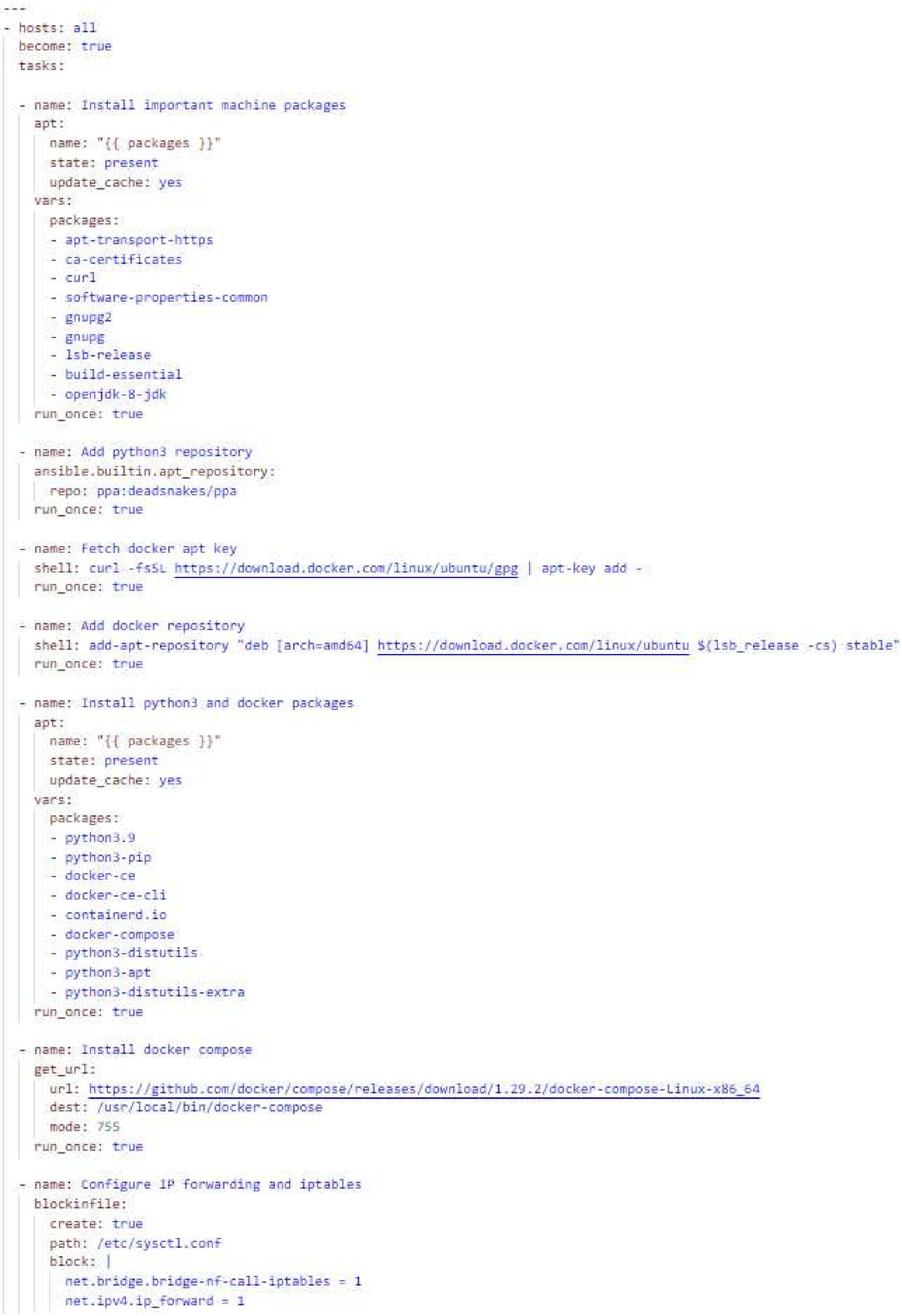
To wrap the execution environment and configuration, Vagrant, Ansible, and Docker give the capacity for *SoDa-TAP* to be deployed on any machine that supports virtualization, in order to automate its deployment as introduced in the beginning of the chapter. This combination of tools tries to solve two problems: (a) to make it easier for developers to replicate environments without worrying about missing dependencies; and (b) to give the platform the configurable flexibility to delimit resources for specific use cases or services. Their interconnected advantage is the benefit of being able to control hardware resources and software rules that are required for the concepts of robustness and scalability of the platform. The following are the services that contribute to these concepts:

- **Kafka:** Its publish/subscribe data processing makes it possible to connect end-to-end data pipelines through its cluster architecture. It is capable to manage data volumes and in the case of experiencing low throughout, it can scale to deploy parallel services to help in the re-balance of the data transfer.

- **Spark:** Its distributed architecture can be useful to manipulate, transform, and apply operations over textual and numerical data with short execution times. To improve processing time, a local deployment can be extended to a cluster environment.

- **CrateDB:** Its highly performant and distributed architecture makes it a database that is capable to work with time-series or text data, with the help of a powerful query language. It has functionalities from the hybrid transaction/analytical processing architecture, since it allows to apply operations over the data right in the database to be used for analytical purposes. It is capable to manage the writing, consultation and storage of different data volumes. Its capability to deploy the database as a cluster in nodes, allows it to scale the deployment into parallel services to help in the re-balance of the data storage.

The current architecture of the platform gives it the flexibility to incorporate containers or services that are easy to be customized through docker compose files, but also, it reduces the time that is needed to deploy and to solve technical issues. For example: (a) the development of a solution can be containerized to automate tests and execution. At the same time, it makes it easy to transfer an application to another machine; and (b) if a container stops working, there is a log functionality to inspect its problem and solve it. The time that it takes to re-deploy is only the time it takes to solve the problem. All services in *SoDa-TAP* have the capability to work in local environments, but also, they are a good fit for clusters of machines. If the case is that performance needs to be upgraded or there are limitations that result in errors, hardware resources can be easily replaced and logical functionalities would not be impacted for such changes.



```
---
- hosts: all
  become: true
  tasks:

  - name: Install important machine packages
    apt:
      name: "{{ packages }}"
      state: present
      update_cache: yes
    vars:
      packages:
      - apt-transport-https
      - ca-certificates
      - curl
      - software-properties-common
      - gnupg2
      - gnupg
      - lsb-release
      - build-essential
      - openjdk-8-jdk
    run_once: true

  - name: Add python3 repository
    ansible.builtin.apt_repository:
      repo: ppa:deadsnakes/ppa
    run_once: true

  - name: Fetch docker apt key
    shell: curl -fsSL https://download.docker.com/linux/ubuntu/gpg | apt-key add -
    run_once: true

  - name: Add docker repository
    shell: add-apt-repository "deb [arch=amd64] https://download.docker.com/linux/ubuntu $(lsb_release -cs) stable"
    run_once: true

  - name: Install python3 and docker packages
    apt:
      name: "{{ packages }}"
      state: present
      update_cache: yes
    vars:
      packages:
      - python3.9
      - python3-pip
      - docker-ce
      - docker-ce-cli
      - containerd.io
      - docker-compose
      - python3-distutils
      - python3-apt
      - python3-distutils-extra
    run_once: true

  - name: Install docker compose
    get_url:
      url: https://github.com/docker/compose/releases/download/1.29.2/docker-compose-Linux-x86_64
      dest: /usr/local/bin/docker-compose
      mode: 755
    run_once: true

  - name: Configure IP forwarding and iptables
    blockinfile:
      create: true
      path: /etc/sysctl.conf
      block: |
        net.bridge.bridge-nf-call-iptables = 1
        net.ipv4.ip_forward = 1
```

Figure 4.2: Part of Ansible Set Up Playbook.

Chapter 5

Our Experience with *SoDa-TAP*

We have applied *SoDa-TAP* in the following case studies. The first was in collaboration with the communications team of the University of Alberta who were interested in better understanding the Twitter conversations about the University. They compiled a list of handles of Twitter accounts associated with units and people on campus and asked us to generate reports on which accounts and what (types of tweets) generated the most/least engagement. The other case study was conducted in collaboration with the Canadian Energy and Climate Nexus organization and was conceived to better understand the Twitter conversations around energy, during the "Energy East Pipeline" timeline. Some of the questions of interest in this case study were the following: how were the people expressing their opinions, what were some of the contexts of those opinions, and what were some of the numerical properties of high engagement opinions. Our visualizations helped us to shed light on these questions, enabling us to explore the data at different time-window granularities and from multiple perspectives. A question of particular interest was "which post elements and features contribute to enhancing its influence?"; hence, we also analyzed content associated with an opinion, in order to correlate its influence to a variety of elements: Twitter's textual elements and media (currently image(s)).

In the next sections, we give more details about each scenario, its objectives, the *SoDa-TAP* components involved in each of them, our findings, and our reflections on the issues that we ran into during development and deployment.

5.1 University of Alberta Engagement on Twitter

SoDa-TAP is currently able to generate three metrics to quantify influence, to bring support to the decision whether an account should be terminated or to pay more attention into content creation.

SoDa-TAP Data Workflow

As one of the earliest studies that we conducted. At the time, only the scraper client was available and this is why it was used to extract the posts of the specific users' accounts collected by the Communications team.

1. **Data Collection.** Individual CSV files were generated from each account and they were aggregated into a single CSV file containing all the tweets of the accounts. In parallel, we generated a single CSV file combining all of the accounts' public metadata such as: followers, number of tweets, etc.

2. **Data Processing and Analysis.** A Spark job consumed the generated CSV file of the accounts' tweets, were it applied all available pipelines: dictionary look up, ABSA and engagement. Once processing was done, the system automatically performed batch writing of the results to database. Following up, a users table was created from the accounts' public metadata CSV file.

3. **Data Storage.** We manually created in CrateDB's web dashboard a view that calculates standard deviation, median, and mean of engagement from all the dataset.

4. **Data Visualization.** A website composed of two tables as previously introduced in Figure 3.7, one for tweets and one for users, were used to allow full dataset exploration of the results.

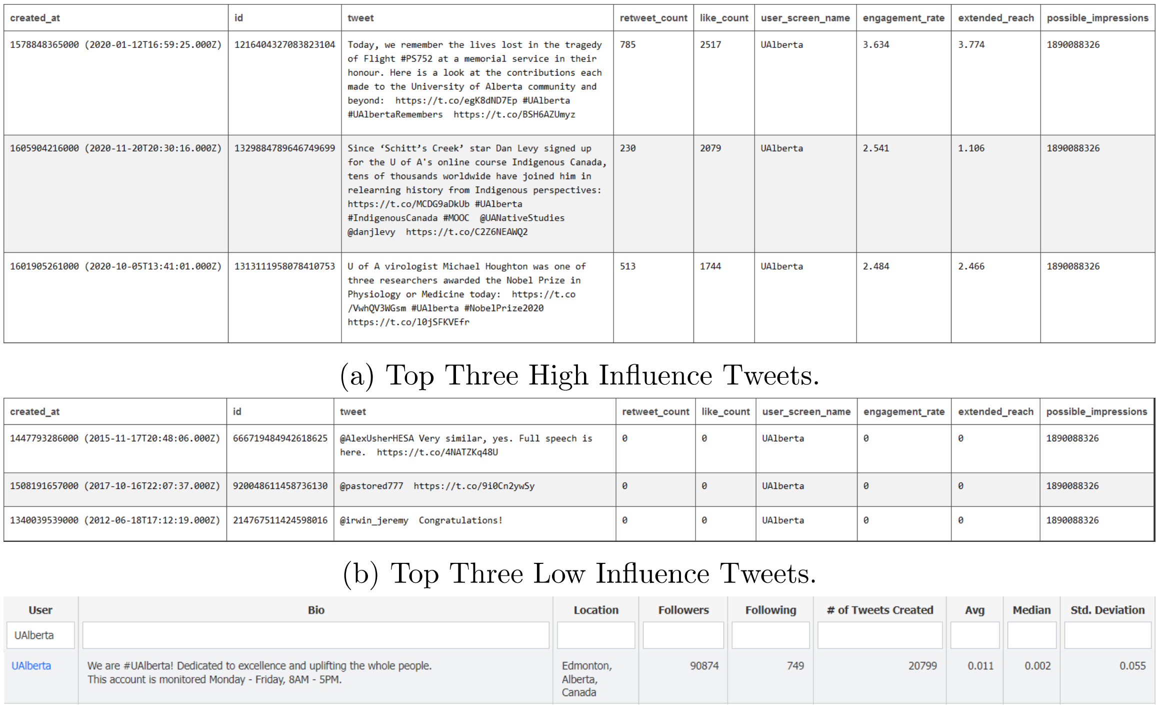
This case study was conducted fairly early in the development of *SoDa-TAP* and, to a degree, motivated some of the subsequent development of the platform. In order to explore the dataset, we primarily took advantage of the table visualizations to study three engagement metrics: (a) engagement rate, which denotes the level of interaction that users have with the post; (b) extended reach, which calculates a ratio of the retweets done from the total of tweets created; and (c) possible impressions, which is a percentage of how many of the followers were reached. The questions below aroused based on the characteristics of this dataset.

How Influential Are the Official UAlberta Accounts?

The first analysis that we did was to study the range of influence metrics of the users of interest. Through the execution of a SQL query as shown in Listing 5.1, we filtered highest and lowest engagement rates from a total of 13,201 tweets. As seen in Figures 5.1a and 5.1b, not surprisingly replies to original tweets generated fewer interactions because of its direct communication to just a single user, compared to a message directed to the general public. The most engaging tweets have high retweets and likes count, while less engaging tweets have zero retweets and likes. The extended reach metric shares the same properties, since its formula relies on likes and retweets as well. In our study, the possible impressions metric is the same for all of the tweets of a user, because of the static number of the user's followers at the time of extraction. In conclusion, the interpretation of these engagement metrics can be that the account is not very effective in attracting users' interactions; as can be seen in Figure 5.1c, the average engagement rate was 0.011, which means that there are a lot of tweets with a null influence. From an account that at the time had a total of 28,799 tweets and 90,874 followers, the potential impressions can be seen high, but in reality the content and users' interactions are the ones that better indicate whether tweets are impactful.

Listing 5.1: SQL Query to Select and Filter Account.

```
SELECT created_at, id, tweet, retweet_count, user_id,
    like_count, user_screen_name, engagement_rate,
    extended_reach, possible_impressions FROM {table} WHERE
    user_screen_name='UAlberta' ORDER BY {public_metric} {DESC
    |ASC};
```



created_at	id	tweet	retweet_count	like_count	user_screen_name	engagement_rate	extended_reach	possible_impressions
1578848365000 (2020-01-12T16:59:25.000Z)	1216404327083823104	Today, we remember the lives lost in the tragedy of Flight #PS752 at a memorial service in their honour. Here is a look at the contributions each made to the University of Alberta community and beyond: https://t.co/egK8dND7Ep #UAlberta #UAlbertaRemembers https://t.co/BSH6AZUmyz	785	2517	UAlberta	3.634	3.774	1890088326
1605904216000 (2020-11-20T20:30:16.000Z)	1329884789646749699	Since 'Schitt's Creek' star Dan Levy signed up for the U of A's online course Indigenous Canada, tens of thousands worldwide have joined him in relearning history from Indigenous perspectives: https://t.co/MCDG9aDkUb #UAlberta #IndigenousCanada #MOOC @UANativeStudies @danjlevy https://t.co/C2Z6NEAWQ2	230	2079	UAlberta	2.541	1.106	1890088326
1601905261000 (2020-10-05T13:41:01.000Z)	1313111958078410753	U of A virologist Michael Houghton was one of three researchers awarded the Nobel Prize in Physiology or Medicine today: https://t.co /VwhQV3WGsm #UAlberta #NobelPrize2020 https://t.co/l0jSFKVEfr	513	1744	UAlberta	2.484	2.466	1890088326

(a) Top Three High Influence Tweets.

created_at	id	tweet	retweet_count	like_count	user_screen_name	engagement_rate	extended_reach	possible_impressions
1447793286000 (2015-11-17T20:48:06.000Z)	666719484942618625	@AlexUsherHESA Very similar, yes. Full speech is here. https://t.co/4NATZKq48U	0	0	UAlberta	0	0	1890088326
1508191657000 (2017-10-16T22:07:37.000Z)	920048611458736130	@pastored777 https://t.co/9i0Cn2ywSy	0	0	UAlberta	0	0	1890088326
1340039539000 (2012-06-18T17:12:19.000Z)	214767511424598016	@irwin_jeremy Congratulations!	0	0	UAlberta	0	0	1890088326

(b) Top Three Low Influence Tweets.

User	Bio	Location	Followers	Following	# of Tweets Created	Avg	Median	Std. Deviation
UAlberta								
UAlberta	We are #UAlberta! Dedicated to excellence and uplifting the whole people. This account is monitored Monday - Friday, 8AM - 5PM.	Edmonton, Alberta, Canada	90874	749	20799	0.011	0.002	0.055

(c) UAlberta Account Metadata.

Figure 5.1: Influence Metrics.

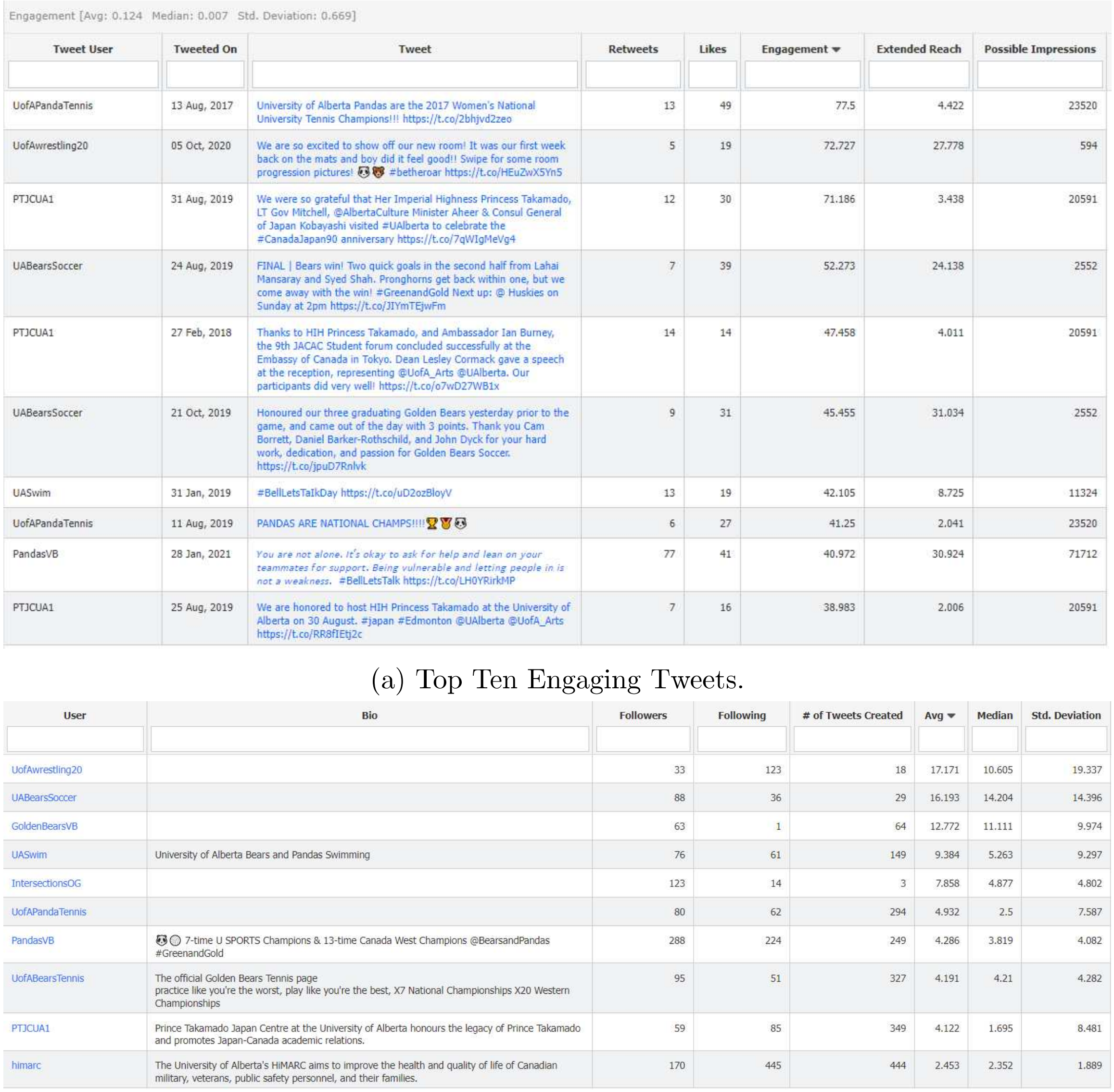
Which Accounts and Tweets Have a High Engagement Rate?

We calculated the engagement distribution of 130 users and their 264,887 tweets: the average was 0.124, the median 0.007 and the standard deviation was 0.669. These results show that in average and the central value for engagement is fairly low, and its relationship with the standard deviation can tell us that even though most of the rates are close to zero, there exists sparsity, considering that the standard deviation is small and close to the mean. The reason of this sparsity could be because of accounts that are more active than other ones, but the most common engagement gets to be low.

Following up, we decided to investigate which ten tweets and ten users were the most engaging as seen in Figure 5.2 and by looking into the metrics, they had high variability. Ranges of engaging tweets varied from approximately 39% from user “PTJCUA1” to 77.5% from user “UofAPandaTennis” and the average of engagement for the top users ranged between 2.4% from user “himarc” to 17.17% from user “UofAWrestling20”. Looking at the intersection between top users and top tweets, the accounts “UofAWrestling20”, “UABearsSoccer”, “UASwim”, “UofAPandaTennis”, “PandasVB”, and “PTJCUA1” had more presence in tweets creation. “PTJCUA1” with 3 tweets, “UofAPandaTennis” and “UABearsSoccer” with 2 tweets, contributed the most to the high interactions for engagement. What we discovered from this analysis was that the followers per account can be relative, one account can have a small audience and have all of them active, which would result in a more engaging content, but also could mean the opposite, an account can have a high audience but they are not getting engaged. To conclude, to tell how well an account is performing, it needs to be visually supported based on the ratio between followers and the interactions that were done based on the number of followers. Engagement rate is highly dependent to followers and interactions, while potential impressions and extended reach to the number of tweets.

5.2 Conversations Around Energy

This case study used the client accessing the Twitter API. The search was based on the hashtag “energyeast”, between the dates March 21, 2006 to June 17, 2021. The resulting dataset was composed of 28,693 tweets, 111,091 retweets, 3,753 tweets with quotes and 4,780 replies with a total of 148,317 tweets, in the period from June 7, 2013 to June 16, 2021. The dataset was first *collected* in a CSV file and in parallel, URL(s) got stored in a text file.



Engagement [Avg: 0.124 Median: 0.007 Std. Deviation: 0.669]

Tweet User	Tweeted On	Tweet	Retweets	Likes	Engagement ▾	Extended Reach	Possible Impressions
UofAPandaTennis	13 Aug, 2017	University of Alberta Pandas are the 2017 Women's National University Tennis Champions!!! https://t.co/2bhjvd2zeo	13	49	77.5	4.422	23520
UofAwrestling20	05 Oct, 2020	We are so excited to show off our new room! It was our first week back on the mats and boy did it feel good!! Swipe for some room progression pictures! 🐼🐻 #betheroar https://t.co/HEuZwX5Yn5	5	19	72.727	27.778	594
PTJCUA1	31 Aug, 2019	We were so grateful that Her Imperial Highness Princess Takamado, LT Gov Mitchell, @AlbertaCulture Minister Aheer & Consul General of Japan Kobayashi visited #UAlberta to celebrate the #CanadaJapan90 anniversary https://t.co/7qWIgMeVg4	12	30	71.186	3.438	20591
UABearsSoccer	24 Aug, 2019	FINAL | Bears win! Two quick goals in the second half from Lahai Mansaray and Syed Shah. Pronghorns get back within one, but we come away with the win! #GreenandGold Next up: @ Huskies on Sunday at 2pm https://t.co/JIYmTEjwFm	7	39	52.273	24.138	2552
PTJCUA1	27 Feb, 2018	Thanks to HIH Princess Takamado, and Ambassador Ian Burney, the 9th JACAC Student forum concluded successfully at the Embassy of Canada in Tokyo. Dean Lesley Cormack gave a speech at the reception, representing @UofA_Arts @UAlberta. Our participants did very well! https://t.co/o7wD27WB1x	14	14	47.458	4.011	20591
UABearsSoccer	21 Oct, 2019	Honoured our three graduating Golden Bears yesterday prior to the game, and came out of the day with 3 points. Thank you Cam Borrett, Daniel Barker-Rothschild, and John Dyck for your hard work, dedication, and passion for Golden Bears Soccer. https://t.co/jpuD7Rnlvk	9	31	45.455	31.034	2552
UASwim	31 Jan, 2019	#BellLetsTaIkDay https://t.co/uD2ozBloyV	13	19	42.105	8.725	11324
UofAPandaTennis	11 Aug, 2019	PANDAS ARE NATIONAL CHAMPS!!!!🏆🏅🐼	6	27	41.25	2.041	23520
PandasVB	28 Jan, 2021	*You are not alone. It's okay to ask for help and lean on your teammates for support. Being vulnerable and letting people in is not a weakness.* #BellLetsTalk https://t.co/LH0YRirkMP	77	41	40.972	30.924	71712
PTJCUA1	25 Aug, 2019	We are honored to host HIH Princess Takamado at the University of Alberta on 30 August. #japan #Edmonton @UAlberta @UofA_Arts https://t.co/RR8fIEtj2c	7	16	38.983	2.006	20591

(a) Top Ten Engaging Tweets.

User	Bio	Followers	Following	# of Tweets Created	Avg ▾	Median	Std. Deviation
UofAwrestling20		33	123	18	17.171	10.605	19.337
UABearsSoccer		88	36	29	16.193	14.204	14.396
GoldenBearsVB		63	1	64	12.772	11.111	9.974
UASwim	University of Alberta Bears and Pandas Swimming	76	61	149	9.384	5.263	9.297
IntersectionsOG		123	14	3	7.858	4.877	4.802
UofAPandaTennis		80	62	294	4.932	2.5	7.587
PandasVB	🐼🏐 7-time U SPORTS Champions & 13-time Canada West Champions @BearsandPandas #GreenandGold	288	224	249	4.286	3.819	4.082
UofABearsTennis	The official Golden Bears Tennis page practice like you're the worst, play like you're the best, X7 National Championships X20 Western Championships	95	51	327	4.191	4.21	4.282
PTJCUA1	Prince Takamado Japan Centre at the University of Alberta honours the legacy of Prince Takamado and promotes Japan-Canada academic relations.	59	85	349	4.122	1.695	8.481
himarc	The University of Alberta's HiMARC aims to improve the health and quality of life of Canadian military, veterans, public safety personnel, and their families.	170	445	444	2.453	2.352	1.889

(b) Top Ten Engaging Accounts.

Figure 5.2: Influence Metrics.

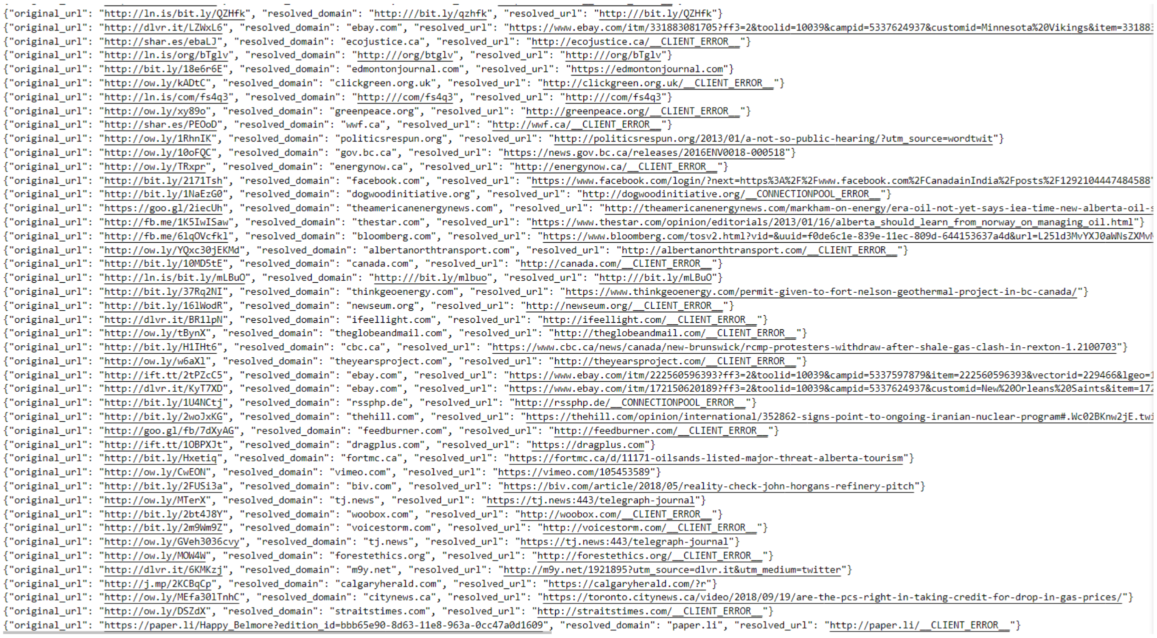
5.2.1 Conversations

The objective of this study was to explore *SoDa-TAP* capabilities to do text-analysis and measure engagement.

SoDa-TAP Data Workflow

One of the first *data processing* operations applied was the expansion of shorten URL(s) with urlExpander [120] from the text files that got created beforehand. The expanded URL(s) would later be saved in a JSON file as seen in Figure 5.3, and subsequently stored in the database.

Next, the ingestion process automatically started and a Spark job was launched to execute a set of pipelines: dictionary look up, ABSA and engagement. After each processing function was done, the system automatically transferred all of the results to the database. Later, a view was created through the execution of a SQL instruction to only show the authors of tweets, as seen in Listing 5.2. To finalize, visualizations from Figures 3.5 and 3.6 were deployed to do an exploratory analysis. A detailed explanation of findings and usability can be found in our paper submitted for "CASCONxEVOKE 2021" conference [123]. The intention of this analysis was to help answer the questions below.



```
{"original_url": "http://ln.is/bit.ly/QZHfk", "resolved_domain": "http:///bit.ly/qzhfk", "resolved_url": "http:///bit.ly/QZHfk"}
{"original_url": "http://dlvr.it/LZWxL6", "resolved_domain": "ebay.com", "resolved_url": "https://www.ebay.com/itm/331883081705?ff3=2&toolid=10039&campid=5337624937&customid=Minnesota%20Vikings&item=33188
{"original_url": "http://shar.es/ebaLJ", "resolved_domain": "ecojustice.ca", "resolved_url": "http://ecojustice.ca/__CLIENT_ERROR__"}
{"original_url": "http://ln.is/org/bTglv", "resolved_domain": "http:///org/btglv", "resolved_url": "http:///org/bTglv"}
{"original_url": "http://bit.ly/18e6r6E", "resolved_domain": "edmontonjournal.com", "resolved_url": "https://edmontonjournal.com"}
{"original_url": "http://ow.ly/kADtC", "resolved_domain": "clickgreen.org.uk", "resolved_url": "http://clickgreen.org.uk/__CLIENT_ERROR__"}
{"original_url": "http://ln.is/com/fs4q3", "resolved_domain": "http:///com/fs4q3", "resolved_url": "http:///com/fs4q3"}
{"original_url": "http://ow.ly/xy89o", "resolved_domain": "greenpeace.org", "resolved_url": "http://greenpeace.org/__CLIENT_ERROR__"}
{"original_url": "http://shar.es/PEOoD", "resolved_domain": "wwf.ca", "resolved_url": "http://wwf.ca/__CLIENT_ERROR__"}
{"original_url": "http://ow.ly/1RhnIK", "resolved_domain": "politicsrespun.org", "resolved_url": "http://politicsrespun.org/2013/01/a-not-so-public-hearing/?utm_source=wordtwit"}
{"original_url": "http://ow.ly/10oFQC", "resolved_domain": "gov.bc.ca", "resolved_url": "https://news.gov.bc.ca/releases/2016ENV0018-000518"}
{"original_url": "http://ow.ly/TRxpr", "resolved_domain": "energynow.ca", "resolved_url": "http://energynow.ca/__CLIENT_ERROR__"}
{"original_url": "http://bit.ly/2171Tsh", "resolved_domain": "facebook.com", "resolved_url": "https://www.facebook.com/login/?next=https%3A%2F%2Fwww.facebook.com%2FCanadainIndia%2Fposts%2F1292104447484588
{"original_url": "http://bit.ly/1NaEzG0", "resolved_domain": "dogwoodinitiative.org", "resolved_url": "http://dogwoodinitiative.org/__CONNECTIONPOOL_ERROR__"}
{"original_url": "https://goo.gl/2iecUh", "resolved_domain": "theamericanenergynews.com", "resolved_url": "http://theamericanenergynews.com/markham-on-energy/era-oil-not-yet-says-iea-time-new-alberta-oil-
{"original_url": "http://fb.me/1K5IwISaw", "resolved_domain": "thestar.com", "resolved_url": "https://www.thestar.com/opinion/editorials/2013/01/16/alberta_should_learn_from_norway_on_managing_oil.html"}
{"original_url": "http://fb.me/6lqOVcfkl", "resolved_domain": "bloomberg.com", "resolved_url": "https://www.bloomberg.com/tosv2.html?vid=&uuid=f0de6c1e-839e-11ec-809d-644153637a4d&url=L251d3MvYXJ0aWNsZXMv
{"original_url": "http://ow.ly/YQxc30jEKMd", "resolved_domain": "albertanorthtransport.com", "resolved_url": "http://albertanorthtransport.com/__CLIENT_ERROR__"}
{"original_url": "http://bit.ly/10MD5tE", "resolved_domain": "canada.com", "resolved_url": "http://canada.com/__CLIENT_ERROR__"}
{"original_url": "http://ln.is/bit.ly/mLBuO", "resolved_domain": "http:///bit.ly/mlbuo", "resolved_url": "http:///bit.ly/mLBuO"}
{"original_url": "http://bit.ly/37Rq2NI", "resolved_domain": "thinkgeoenergy.com", "resolved_url": "https://www.thinkgeoenergy.com/permit-given-to-fort-nelson-geothermal-project-in-bc-canada/"}
{"original_url": "http://bit.ly/16lWodR", "resolved_domain": "newseum.org", "resolved_url": "http://newseum.org/__CLIENT_ERROR__"}
{"original_url": "http://dlvr.it/BR1lpN", "resolved_domain": "ifeellight.com", "resolved_url": "http://ifeellight.com/__CLIENT_ERROR__"}
{"original_url": "http://ow.ly/tBynX", "resolved_domain": "theglobeandmail.com", "resolved_url": "http://theglobeandmail.com/__CLIENT_ERROR__"}
{"original_url": "http://bit.ly/H1IHt6", "resolved_domain": "cbc.ca", "resolved_url": "https://www.cbc.ca/news/canada/new-brunswick/rcmp-protesters-withdraw-after-shale-gas-clash-in-rexton-1.2100703"}
{"original_url": "http://ow.ly/w6aXl", "resolved_domain": "theyearsproject.com", "resolved_url": "http://theyearsproject.com/__CLIENT_ERROR__"}
{"original_url": "http://ift.tt/2tPZcC5", "resolved_domain": "ebay.com", "resolved_url": "https://www.ebay.com/itm/222560596393?ff3=2&toolid=10039&campid=5337597879&item=222560596393&vectorid=229466&lgeo=
{"original_url": "http://dlvr.it/KyT7XD", "resolved_domain": "ebay.com", "resolved_url": "https://www.ebay.com/itm/172150620189?ff3=2&toolid=10039&campid=5337624937&customid=New%20Orleans%20Saints&item=172
{"original_url": "http://bit.ly/1U4NCtj", "resolved_domain": "rssphp.de", "resolved_url": "http://rssphp.de/__CONNECTIONPOOL_ERROR__"}
{"original_url": "http://bit.ly/2woJxKG", "resolved_domain": "thehill.com", "resolved_url": "https://thehill.com/opinion/international/352862-signs-point-to-ongoing-iranian-nuclear-program#.Wc02BKnw2jE.twi
{"original_url": "http://goo.gl/fb/7dXyAG", "resolved_domain": "feedburner.com", "resolved_url": "http://feedburner.com/__CLIENT_ERROR__"}
{"original_url": "http://ift.tt/1OBPXJt", "resolved_domain": "dragplus.com", "resolved_url": "https://dragplus.com"}
{"original_url": "http://bit.ly/Hxetiq", "resolved_domain": "fortmc.ca", "resolved_url": "https://fortmc.ca/d/11171-oilsands-listed-major-threat-alberta-tourism"}
{"original_url": "http://ow.ly/CwEON", "resolved_domain": "vimeo.com", "resolved_url": "https://vimeo.com/105453589"}
{"original_url": "http://bit.ly/2FUSi3a", "resolved_domain": "biv.com", "resolved_url": "https://biv.com/article/2018/05/reality-check-john-horgans-refinery-pitch"}
{"original_url": "http://ow.ly/MTerX", "resolved_domain": "tj.news", "resolved_url": "https://tj.news:443/telegraph-journal"}
{"original_url": "http://bit.ly/2bt4J8Y", "resolved_domain": "woobox.com", "resolved_url": "http://woobox.com/__CLIENT_ERROR__"}
{"original_url": "http://bit.ly/2m9Wm9Z", "resolved_domain": "voicestorm.com", "resolved_url": "http://voicestorm.com/__CLIENT_ERROR__"}
{"original_url": "http://ow.ly/GVeh3036cvy", "resolved_domain": "tj.news", "resolved_url": "https://tj.news:443/telegraph-journal"}
{"original_url": "http://ow.ly/MOW4W", "resolved_domain": "forestethics.org", "resolved_url": "http://forestethics.org/__CLIENT_ERROR__"}
{"original_url": "http://dlvr.it/6KMKzj", "resolved_domain": "m9y.net", "resolved_url": "http://m9y.net/1921895?utm_source=dlvr.it&utm_medium=twitter"}
{"original_url": "http://j.mp/2KCBqCp", "resolved_domain": "calgaryherald.com", "resolved_url": "https://calgaryherald.com/?r"}
{"original_url": "http://ow.ly/MEfa30lTnhC", "resolved_domain": "citynews.ca", "resolved_url": "https://toronto.citynews.ca/video/2018/09/19/are-the-pcs-right-in-taking-credit-for-drop-in-gas-prices/"}
{"original_url": "http://ow.ly/DSZdX", "resolved_domain": "straitstimes.com", "resolved_url": "http://straitstimes.com/__CLIENT_ERROR__"}
{"original_url": "https://paper.li/Happy_Belmore?edition_id=bbb65e90-8d63-11e8-963a-0cc47a0d1609", "resolved_domain": "paper.li", "resolved_url": "http://paper.li/__CLIENT_ERROR__"}
```

Figure 5.3: Example of Expanded URLs JSON File.

Listing 5.2: SQL Query to Select Author of Tweet.

```
SELECT DISTINCT author_id, author_username, author_bio,
    author_followers_count, author_following_count,
    author_tweet_count, max(created_at) FROM {table} GROUP BY
    author_id, author_username, author_bio,
    author_followers_count, author_following_count,
    author_tweet_count;
```

Which Date Had the Most Volume of Conversations and What Were People's Personal Values and Sentiment Dominant Contexts?

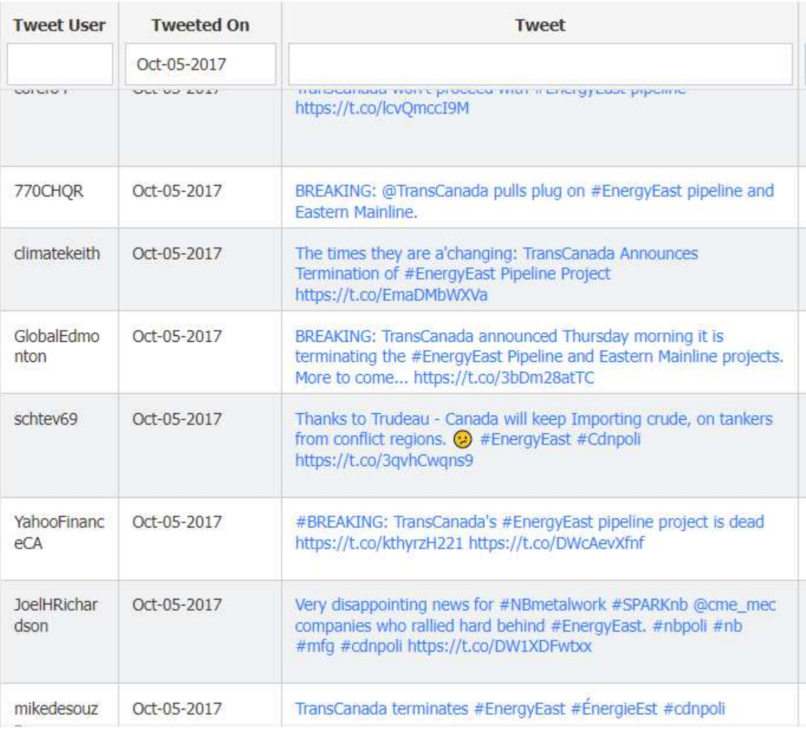
To help answer this question, we first picked the "line and area chart" visualization were we detected four high peaks, independently of the dictionary selected: August 16, 2014; August 12, 2015; January 27, 2016; and October 05, 2017. October 05, 2017 had, by far, the highest volume of conversations, which we selected as our date of study. In this date, the energy east pipeline got cancelled, and this event generated a total of 977 tweets, replies and tweets with quotes. After our data analysis, we appreciated that: (a) the top three personal values found were achievement with 222 words, self direction with 181 words and power with 173 words from a total of 1,142 personal values words found; and (b) there were more negative words than positive, from a total of 1,664 words, 886 were negative and 778 were positive.

What Were Some of the Energy Related Aspects and Their Sentiment?

To analyze in more detail energy-related words, we selected the "treemap chart" for the discovery of personal values and aspect related terms, to get more granular exploration of the words, since it allows to visualize the words label used and frequency. By selecting the date October 05, 2017, we were able to learn that: (i) from the top personal values words, job with 34 words, decision with 50 words and victory with 33 words were the most frequently used; and (ii) aspects related to employment, cost and sustainability with a neutral sentiment were mostly used. These results, guided us to infer that people's opinions were fairly balanced between supporters and opponents. Figure 5.4 shows some of the tweets that were shared in this timeline.

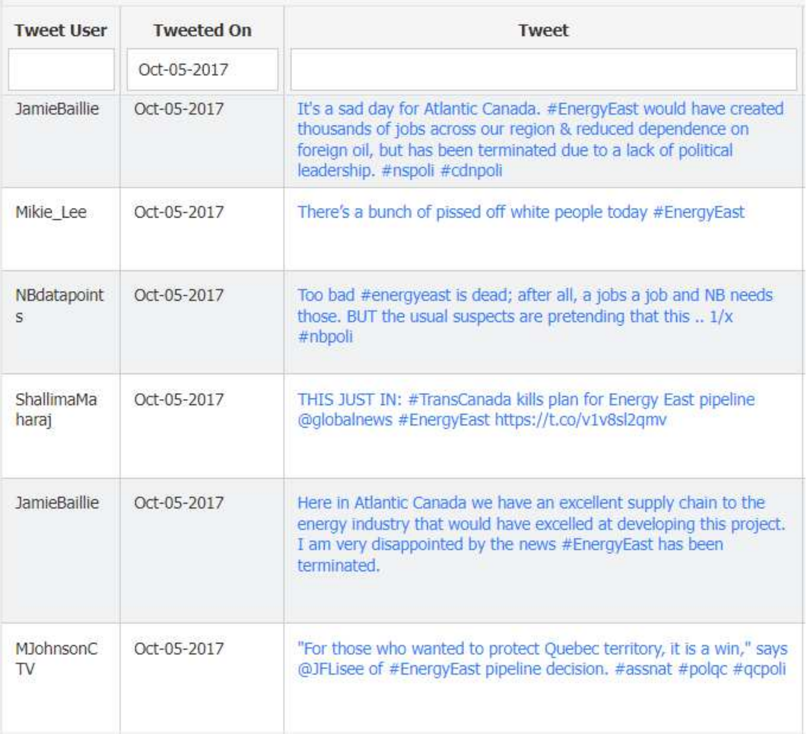
Which Tweets Had the Highest Engagement Rate?

From this event, using the "table" visualization, we observed that 629 users participated in the conversations. The top five tweets that had the highest engagement rate, as seen in Figure 5.5, had a high range difference, from 9.66% to 54.54%. Based on



Tweet User	Tweeted On	Tweet
	Oct-05-2017	
[illegible]	[illegible]	TransCanada won't proceed with #EnergyEast pipeline https://t.co/lcvQmccI9M
770CHQR	Oct-05-2017	BREAKING: @TransCanada pulls plug on #EnergyEast pipeline and Eastern Mainline.
climatekeith	Oct-05-2017	The times they are a'changing: TransCanada Announces Termination of #EnergyEast Pipeline Project https://t.co/EmaDMbWXVa
GlobalEdmonton	Oct-05-2017	BREAKING: TransCanada announced Thursday morning it is terminating the #EnergyEast Pipeline and Eastern Mainline projects. More to come... https://t.co/3bDm28atTC
schtev69	Oct-05-2017	Thanks to Trudeau - Canada will keep Importing crude, on tankers from conflict regions. ☹ #EnergyEast #Cdnpoli https://t.co/3qvhCwqns9
YahooFinanceCA	Oct-05-2017	#BREAKING: TransCanada's #EnergyEast pipeline project is dead https://t.co/kthyrzH221 https://t.co/DWcAevXfnf
JoelHRichardson	Oct-05-2017	Very disappointing news for #NBmetalwork #SPARKnb @cme_mec companies who rallied hard behind #EnergyEast. #nbpoli #nb #mfg #cdnpoli https://t.co/DW1XDFwtxx
mikedesouz	Oct-05-2017	TransCanada terminates #EnergyEast #ÉnergieEst #cdnpoli

(a) Energy East Tweets.

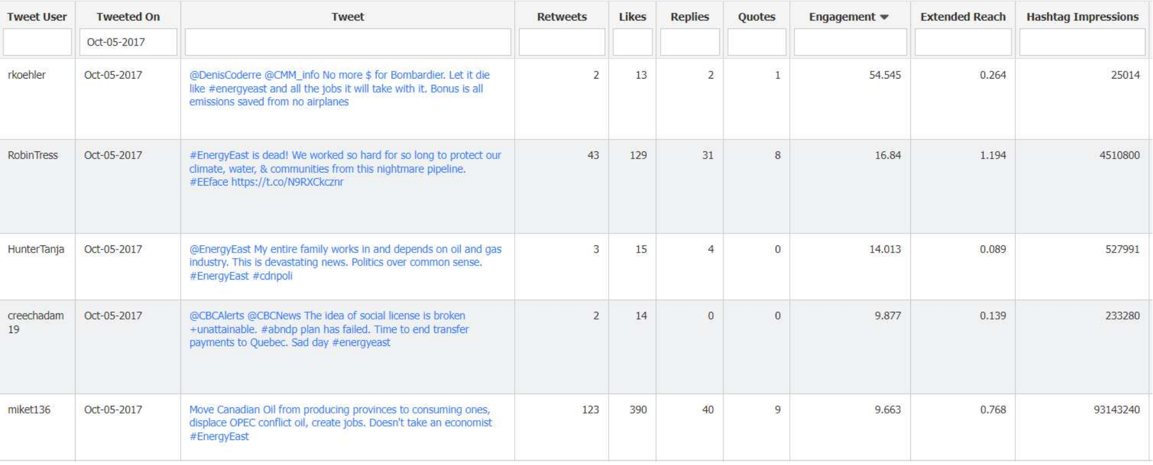


Tweet User	Tweeted On	Tweet
	Oct-05-2017	
JamieBaillie	Oct-05-2017	It's a sad day for Atlantic Canada. #EnergyEast would have created thousands of jobs across our region & reduced dependence on foreign oil, but has been terminated due to a lack of political leadership. #nspoli #cdnpoli
Mikie_Lee	Oct-05-2017	There's a bunch of pissed off white people today #EnergyEast
NBdatapoints	Oct-05-2017	Too bad #energyeast is dead; after all, a jobs a job and NB needs those. BUT the usual suspects are pretending that this .. 1/x #nbpoli
ShallimaMaharaj	Oct-05-2017	THIS JUST IN: #TransCanada kills plan for Energy East pipeline @globalnews #EnergyEast https://t.co/v1v8sl2qmv
JamieBaillie	Oct-05-2017	Here in Atlantic Canada we have an excellent supply chain to the energy industry that would have excelled at developing this project. I am very disappointed by the news #EnergyEast has been terminated.
MJohnsonCTV	Oct-05-2017	"For those who wanted to protect Quebec territory, it is a win," says @JFLisee of #EnergyEast pipeline decision. #assnat #polqc #qcpoli

(b) Energy East Tweets.

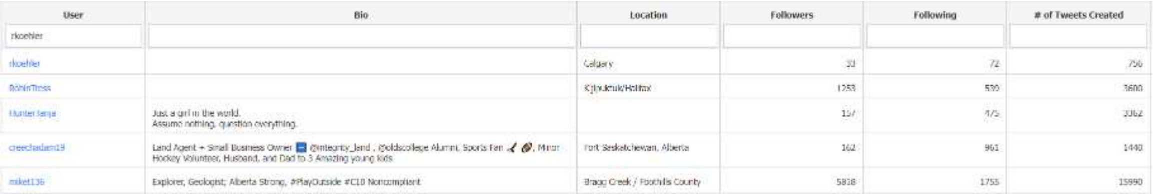
Figure 5.4: Energy East Tweets Sample.

the tweets' likes, replies and quotes, we observed why such a difference. The tweets correlated to their respective author's metadata; as seen in Figure 5.6, the ratio of followers and their public metrics make an impact for influence metrics as we found out in our previous analysis.



Tweet User	Tweeted On	Tweet	Retweets	Likes	Replies	Quotes	Engagement ▼	Extended Reach	Hashtag Impressions
	Oct-05-2017								
rkoehler	Oct-05-2017	@DenisCoderre @CMM_info No more $ for Bombardier. Let it die like #energyeast and all the jobs it will take with it. Bonus is all emissions saved from no airplanes	2	13	2	1	54.545	0.264	25014
RobinTress	Oct-05-2017	#EnergyEast is dead! We worked so hard for so long to protect our climate, water, & communities from this nightmare pipeline. #EEface https://t.co/N9RXCkcznr	43	129	31	8	16.84	1.194	4510800
HunterTanja	Oct-05-2017	@EnergyEast My entire family works in and depends on oil and gas industry. This is devastating news. Politics over common sense. #EnergyEast #cdnpoli	3	15	4	0	14.013	0.089	527991
creechadam19	Oct-05-2017	@CBCAlerts @CBCNews The idea of social license is broken +unattainable. #abndp plan has failed. Time to end transfer payments to Quebec. Sad day #energyeast	2	14	0	0	9.877	0.139	233280
miket136	Oct-05-2017	Move Canadian Oil from producing provinces to consuming ones, displace OPEC conflict oil, create jobs. Doesn't take an economist #EnergyEast	123	390	40	9	9.663	0.768	93143240

Figure 5.5: Top Five Energy East Tweets with High Engagement Rate.



User	Bio	Location	Followers	Following	# of Tweets Created
rkoehler					
rkoehler		Calgary	33	72	756
RobinTress		Kjipuktuk/Halifax	1253	539	3600
HunterTanja	Just a girl in the world. Assume nothing, question everything.		157	475	3362
creechadam19	Land Agent + Small Business Owner 🟦 @integrity_land , @oldscollege Alumni, Sports Fan 🏒 🏈, Minor Hockey Volunteer, Husband, and Dad to 3 Amazing young kids	Fort Saskatchewan, Alberta	162	961	1440
miket136	Explorer, Geologist; Alberta Strong, #PlayOutside #C10 Noncompliant	Bragg Creek / Foothills County	5818	1755	15990

Figure 5.6: Top Five Energy East Tweets' Author.

Through this case study, we demonstrated that the visual components of *SoDa-TAP* can be used to support visual data exploration and discovery of significant events and important dates. At the same time, the text-analysis functions are really helpful for revealing insights around the context of the conversations.

5.2.2 Content Engagement

One of the features that *SoDa-TAP* contributes for in-depth data analysis, is its statistical component that allows to keep expanding for custom functions. Under this same concept, we formulated a set of content multimodality and complexity

Table 5.1: Energy Groups and Processing Wall Time.

Group (Energy)	**Size (No. of Tweets)**	**Processing Wall Time (Seconds)**
Carbon Capture Storage	4,664	51
Bio Mass & Biofuels	6,844	60
Coal	7,288	50
Hydrocarbon Gas Liquids	10,722	58
Hydrogen	10,948	59
Geothermal	17,186	66
Refinery	34,059	96
Ocean	38,908	180
Nuclear	134,685	195
Thermal Electric	156,575	210
Wind	202,850	275
Solar	341,331	381
Hydro Electric	371,981	432
Oil Sands	547,195	584
Oil & Gas	1,805,570	1,938

indicators, that were used to explore characteristics of Twitter posts in relationship with their engagement metric.

In order to evaluate if a tweet had media, we had to rehydrate the previously created dataset, to get the image object, when present in a tweet. This new dataset was composed of 28,319 tweets, 107,867 retweets, 3,695 tweets with a quote and 4,680 replies with a total of 144,561 tweets, in the period from June 7, 2013 to June 16, 2021 with a count of 5,696 found images. This dataset was later saved as a single CSV file containing all the required data for processing.

SoDa-TAP Data Workflow

We initialized a Jupyter notebook with the new energy east dataset loaded. The main set of functions that we developed to do *processing and analysis* of this dataset were: (i) modality index, which is a number that ranges from 1 to 6 (plus one for each new element: text, URLs, hashtags, handles, emojis and images); (ii) complexity index, which is a number that ranges from 1 to 4 to define combinations such as text only, text and other elements, text and images, and text, images and other elements, respectively; and (iii) quartiles based on text characters (Q1: $\leq$ 104, Q2: >104 and $\leq$ 117, Q3: >117 and $\leq$ 136, and Q4: >136). At the same time, we started some preliminary work with the image analysis pipeline that applied all of the available functions to the dataset images. To review the results of the analyses, we made use of the box plot *visualization* to show the different variations and correlations that our custom defined indexes presented alongside the public metrics.

The purpose of this demonstration was to interact with the available functions and components from the analysis toolkit to discover insights about concepts around influence of tweets based on visual characteristics, more than just public metrics. The previous influence demonstration and further exploratory analysis to comprehend a tweet's engagement awaken the questions below.

What is the Modality of a Tweet and What Interaction Impacts the Most its Influence?

To evaluate different interactions mechanics available in a tweet, we performed a modality index analysis based on the indexes 1, 2, 3, 4, 5 or 6, explained before. The results, as shown in Table 5.3, indicate that there were three predominant indexes: 4, 2 and 3. Where 3 contained the highest number of tweets, which means that the presence of at least three elements was mostly used. The correlation of this index with their public metrics showed us that "likes‘' and "retweets‘' generated the most impact to make a tweet reach more public. This behaviour could be because of the easier accessibility that a user can whether press a “like” or “retweet” button.

What is the Impact of the Length of a Tweet with its Influence?

We analyzed the relationship between the tweet length divided in groups of quartiles and the public metrics only from original tweets, as seen in Figure 5.7. We observed that “likes” and “retweets” had the most interactions and more specifically these interactions happened the most in tweets that contained a length between 104 and 117 characters and longer than 136. This lead us to infer that in the case of “likes‘’, users get more engaged because of two reasons: (i) the message can potentially contain only text and be not too short; and (ii) the message can potentially contain text plus several URLs, which can impact the character count. And in the case of “retweets‘’, users get more engaged while the message is also not too short.

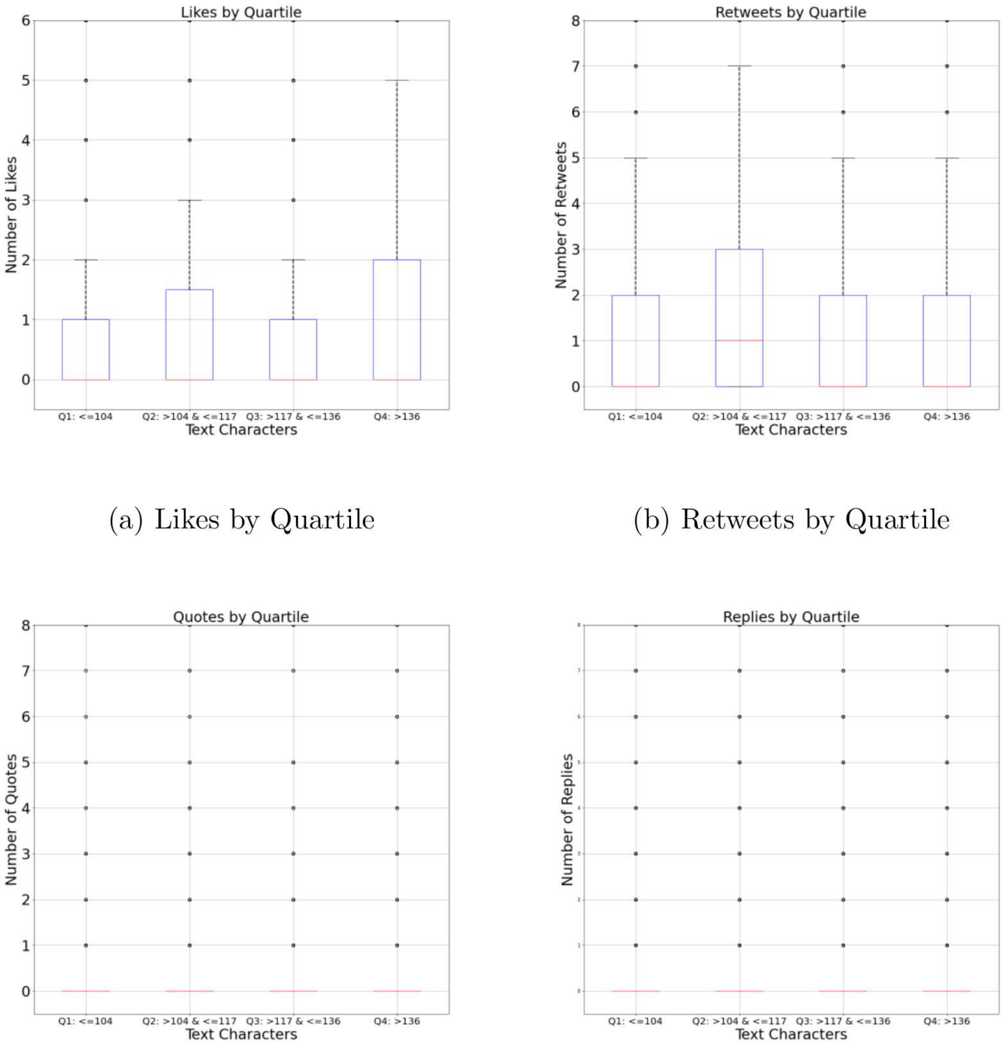
What is the Impact of the Modality of a Tweet with its Influence?

As shown in Figure 5.8, the modality indexes that had the most sparsity were the indexes of 4, 5 and 6. In the case of “likes”, there is an increase of interactions at indexes 5 and 6. In “retweets”, indexes 4, 5 and 6 share almost same increase. For “quotes”, index 6 is the only one with interactions and for “replies”, index 6 shows a significant increase over index 5. At the same time, we observed that tweets that contain mostly one or two of the elements do not tend to receive as many reactions compared to tweets that are constituted with multiple elements.

What is the Impact of the Complexity of a Tweet with its Influence?

As shown in Figure 5.9, tweets that had an image associated were more likely to generate a reaction, except for the “quotes”. One of the reason of this behaviour is because visual elements are more attractive and engaging to the eye than simply text. An observed reason of why “quotes” is missing reactions, could be because the action of quoting a tweet is more complicated than the other set of functionalities available for a tweet. One thing that we noticed is that the variability that the elements add to the interactions is almost insignificant, compared to when a picture is present.

To conclude, these set of early results pointed us that tweets that contain a combination of visuals, whether is an image or Twitter's elements, they attract the user to get involved with a reaction, more frequently through a 'like" or "retweet" because it is easier than trying to compose a complicated reply or tweet.



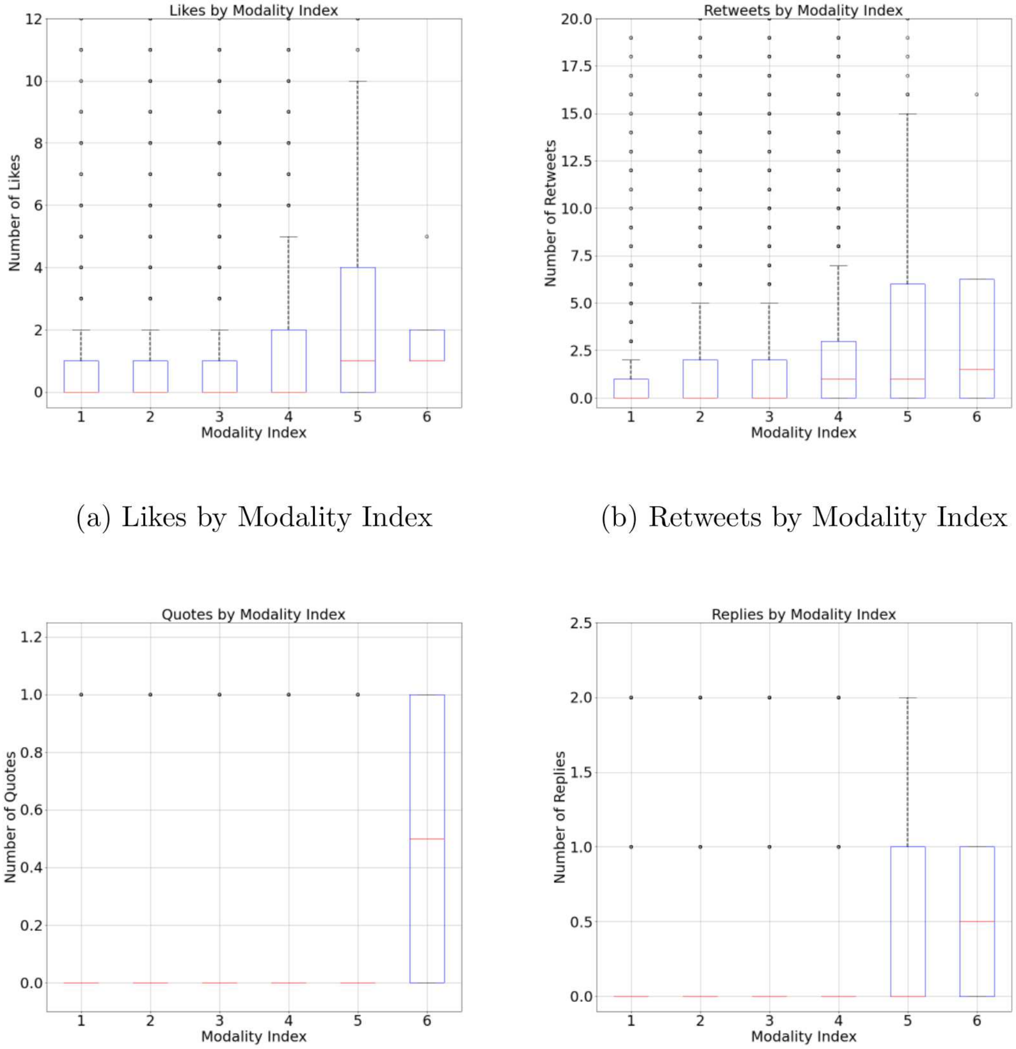
(a) Likes by Quartile

(b) Retweets by Quartile

(c) Quotes by Quartile

(d) Replies by Quartile

Figure 5.7: Tweet Length and Public Metrics.



(a) Likes by Modality Index

(b) Retweets by Modality Index

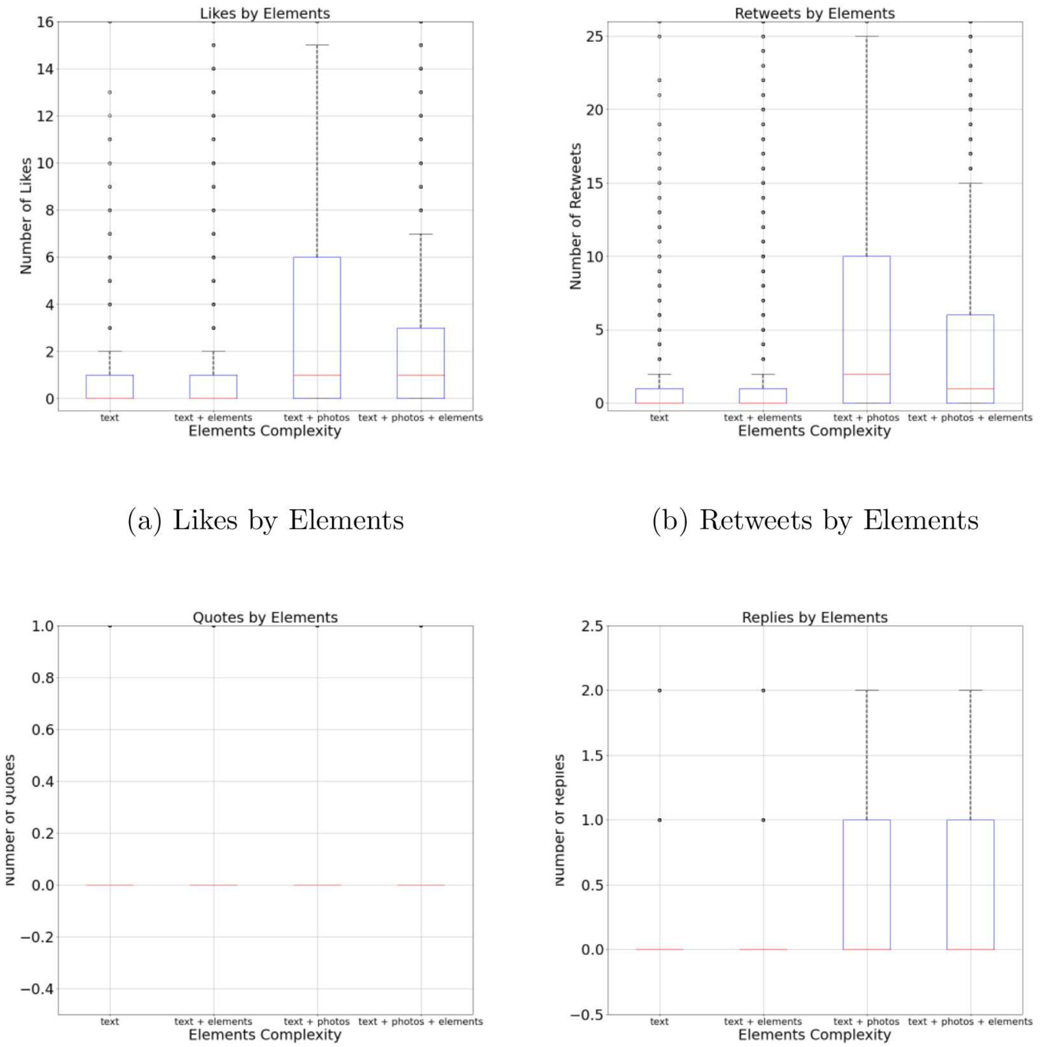
(c) Quotes by Modality Index

(d) Replies by Modality Index

Figure 5.8: Tweet Modality and Public Metrics.

5.2.3 Data Processing Performance

To more accurately analyze how much time data processing takes, we measured the time required for the text and image analysis components that were deployed in different servers. In the first server, composed of a Linux PC with 16 cores and 56 GB of RAM, was used to do mainly text processing. On another server, composed of a Linux PC with 4 cores, 22 GB of RAM and an NVIDIA vGPU with 8 GB of memory, was used to do mainly image processing.



(a) Likes by Elements

(b) Retweets by Elements

(c) Quotes by Elements

(d) Replies by Elements

Figure 5.9: Tweet Elements and Public Metrics.

One of the earliest performance measurements that we conducted was based on the "Energy East" dataset composed of 148,317 tweets, to measure the processing involved in all text-analysis functions. This dataset took approximately 2 minutes between processing and writing to the database.

As part of our exploratory performance measurement, we replicated the *SoDa-TAP* pipelines through fifteen different deployments. Each deployment's dataset was first constructed based on individual queries search through the official API that were related to energy types. As seen in Table 5.2, each energy type had a set of keywords, individually searched. We performed an aggregation and discarded repeated tweets per energy group, to later save each one of the groups as a CSV file. After the creation of each individual dataset and aggregation, *SoDa-TAP* performed all text-analysis pipelines. Preliminary results of the processing performance for these datasets can be seen in Table 5.1. At first sight, we noticed that the largest dataset "Oil & Gas" composed of 1,805,570 tweets was the most computationally expensive with a wall time of approximately 1,938 seconds (32.3 minutes) and the best performant was "Carbon Capture Storage" composed of 4,664 tweets, with a wall time of 51 seconds (0.85 minutes). From these processing times, it can be inferred that for each dataset, each approximately 100,000 records adds up to degrading the processing time. The processing times of the pipelines are analogous to the dataset's size but the relationship is sublinear, which provides evidence that the pipeline scales well as the dataset size increases.

Later, we decided to measure the image analysis processing from the 5,696 images that we got from rehydration. The wall time that we got from applying all the image processing functions took approximately 68,400 seconds (19 hours). Next, we applied the same amount of processing functions over the 68,350 images that we got from the largest dataset. The processing time that it took was of approximately 864,227 seconds (10 days).

To conclude, a possible reason for the variabilities of text and image analysis processing pipelines is because of limited computing resources. Even though they differed in execution time depending on the data size, we did not experience any resource consumption problems during deployment and usability.

5.3 Development and Deployment Challenges

The development and deployment of a robust and large-scale system like *SoDa-TAP* raises many challenges that need to be taken care of and closely monitored.

These concerns are related to components' errors, computing resources limitations, security and privacy. With respect to component's errors, we encountered that individual components problems are usually linked to the computing resources limitations. The components should be deployed in an adequate environment, where resources can be easily increased or decreased, depending on the workload that will be executing. For example, a cloud service could possibly improve the overall execution by giving more freedom to increase local storage and RAM memory.

Network connectivity issues also form part of this set of problems, since our system involve operations that need to be done through the network. Part of the solution to this problem is already being treated by the selected set of tools and components that internally have instructions to deal with this kind of errors, but we can not control external services that are not part of *SoDa-TAP* , which we did not fully explore because it is fully dependable on the environment that it is deployed. That is why we make sure that all of the configuration options for our components are the most generic, independently if it is launched locally or in the cloud, since we know how important it is fault tolerance and scalability while the components are working together.

Security and privacy are two serious topics that need special care because of government regulations that protect users' identity online and to protect each system's component against malicious third-parties. *SoDa-TAP* currently complies in both scenarios by applying the minimum requirements that allow only our components to be able to communicate with each other and the data generated by the system is at all times following the accessibility standards.

Clearly, there are many research opportunities in a system like this, that ranges from the adaptability and capability to support a variety of social media data studies, to the extent to evaluate individual aspects from each of the components. These would contribute to the development of more reliable and robust components, to help with the standardization of social data processing and analysis.

Table 5.2: Energy Type and Keywords.

Energy Type	**Keywords**
Biomass/Biofuel	biofuels, biomass waste, manufacturing waste, biomass, feedstock, bioethanol, landfill gas, biowaste, municipal waste, sugar cane, renewable organic material, agricultural waste.
Carbon Capture Storage	carbon capture storage, CCUS, carbon capture use and storage, coLo, carbon capture utilization and storage, noCCS, carbon capture and storage, carbon capture.
Coal	peat, coker unit, bituminous coal, anthracite, graphite.
Geothermal	geothermal, heat pump, hydrothermal.
Hydrocarbon Gas Liquids	ethane, natural gas liquids, propane, NGL, HGL, liquified natural gas, butane.
Hydro Electric	hydro, hydroelectric, dam, hydropower, run-of-river.
Hydrogen	hydrogen, blue hydrogen, hydrogen fuel, blue H2, carbon-free fuel, green hydrogen, hydrogen energy, green H2, H2.
Nuclear	nuclear, isotope, radiation, fission, reactor, uranium, CANDU, neutron, radioactive, nuclear fusion, nuclear energy.
Ocean Energy	ocean energy, surface waves, tides, wave power, tidal power, marine energy, salinity, hydrokinetic energy, ocean power, marine power, ocean waves, ocean temperature.
Oil & Gas	bitumen, crude oil, dilbit, ethical oil, fossil fuel, fracking, gas, hydrocarbon gas liquids, hydrocarbon, natural gas, oil, petroleum, oil well, hydraulic fracturing, oil reserve.
Oil Sands	oil sands, oilsands, tailings ponds, tar sands, tarsands, tailings, SAGD.
Refinery	refineries, petrochemical, upgrader, refinery.
Solar	solar photovoltaic, solar, solar pv, solar energy, Feed In Tariff (FIT), Power Purchase Agreement, solar panel, Power Purchase Agreement (PPA), Feed In Tariff.
Thermal Electric	coal, open pit, thermal electric, coal mine, coal fired, gas fired, power plant, co-gen, powerplant, coal power plant.
Wind	wind, wind farm, wind energy, wind turbine.

Table 5.3: Multimodality and Public Metrics Descriptive Statistics.

Index	No. of Tweets	Public Metrics															
		Likes				Retweets				Quotes				Replies			
		min	**max**	**avg**	**med**	**min**	**max**	**avg**	**med**	**min**	**max**	**avg**	**med**	**min**	**max**	**avg**	**med**
1	1,659	0	1,249	4.250	0	0	952	3.455	0	0	103	0.246	0	0	226	0.630	0
2	7,592	0	2,506	4.691	0	0	1,291	4.203	0	0	116	0.236	0	0	218	0.642	0
3	12,605	0	1,090	2.753	0	0	492	3.045	0	0	72	0.139	0	0	142	0.352	0
4	5,978	0	1,484	3.465	0	0	555	4.320	1	0	83	0.176	0	0	280	0.456	0
5	465	0	237	6.191	1	0	181	7.374	1	0	20	0.267	0	0	20	0.791	0
6	4	1	5	2	1	0	16	4.750	1.5	0	1	0.5	0.5	0	1	0.5	0.5

Chapter 6

Conclusions and Future Work

In this thesis, we have introduced *SoDa-TAP* , a system for social media data processing and analyses. A variety of studies can be supported by *SoDa-TAP* ; new methods for data processing are easy to be integrated into the system; statistical analyses can be easily extended to support custom evaluations; and the visualizations catalog can be expanded and customized to support a variety of insights discovery. These sets of characteristics separate and differentiate *SoDa-TAP* from existing research and applications. Its automatic deployment capability makes it interesting to researchers and less-technical users.

This thesis makes the following contributions:

1. **The implementation of the *SoDa-TAP* system.** The system has been implemented in a microservices and container based style, which makes all components of the system modular, allowing a drag and drop functionality for functionality extension. A custom configuration file has been designed for automatic deployment of the system. Two API services have been developed to allow for data access, for either a continuous or single data request. A set of custom designed visualizations have been developed to help in the insight discovery. A scalable statistical analysis component has been designed to allow custom evaluation if required.

2. **The development of custom pipelines.** A standard set of processing and

analysis pipelines have been developed to process text, images and URLs. A set of functions have been developed to evaluate multi-modality, complexity, and correlation with public metrics.

3. **Preliminary results from custom analyses to demonstrate the system's adaptability.** Our demonstrations and early results support our statement that the system can be applied in a variety of studies. The "Engagement" analyses demonstrates the capability of the system to measure influence. The "Conversations and Content Engagement" analyses demonstrate the usefulness of all the components of the platform and our statistical analysis component to design custom evaluations of data.

6.1 Future Work

We plan to further develop the system to make it more production-ready by deploying it in a more robust server and expanding it to be used in a cluster of machines, so that further scalability can be feasible. At the same time, we intend to integrate a new set of functionalities as described below.

To support social-network analysis, we will integrate the calculation of centrality measures, among the authors and tweets based on all available relationships of metadata and public metrics: "follows", "likes", "replies", "retweets". Also, a "cluster of influence" will be developed for the analysis and visualization of organizations and users.

The statistical analysis component will be extended to support MANOVA and many more test types - Tukey HSD, Shapiro Wilk, Normal Q-Q plot, Levene, t-test - to be evaluated over selected data subsets based on independent (features of posts) and dependent (influence) variables. Furthermore, more visualizations will be incorporated to support the interpretation of such evaluations.

A real-time data processing component will be integrated to the processing pipeline, composed of a streaming client to fetch posts being created as a stream, and a SQL-like query interface to inspect and apply query operations over incoming data.

Finally, more dictionaries will be integrated to support more text analysis.

In parallel, we will be updating and optimizing the already developed text, images and URLs pipelines for a better processing time, and integrate more methodologies for analysis, including ML models.

SOCMINT is a very broad domain that can have many use cases, but the most important of them all is its use for social causes, to learn from the community's generated data and apply it to improve aspects of the society. We believe that *SoDa-TAP* bridges this gap and makes the improved social media monitoring for the wellness of the community possible.

References

[1] E. Şuşnea and A. Iftene, "The significance of online monitoring activities for the social media intelligence (socmint)," in *Conference on Mathematical Foundations of Informatics*, 2018, pp. 230–240.

[2] R. Dover, "Socmint: A shifting balance of opportunity," *Intelligence and National Security*, vol. 35, no. 2, pp. 216–232, 2020.

[3] B. Sverdrup-Thygeson and V. Engesæth, "Open-source and social media intelligence," in *Intelligence Analysis in the Digital Age*, Routledge, 2021, pp. 52–67.

[4] I. Lee, "Social media analytics for enterprises: Typology, methods, and processes," *Business Horizons*, vol. 61, no. 2, pp. 199–210, 2018.

[5] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[6] H. Sebei, M. A. Hadj Taieb, and M. Ben Aouicha, "Review of social media analytics process and big data pipeline," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–28, 2018.

[7] S. Stieglitz, L. Dang-Xuan, A. Bruns, and C. Neuberger, "Social media analytics," *Business & Information Systems Engineering*, vol. 6, no. 2, pp. 89–96, 2014.

[8] B. Batrinca and P. C. Treleaven, "Social media analytics: A survey of techniques, tools and platforms," *Ai & Society*, vol. 30, no. 1, pp. 89–116, 2015.

[9] Facebook, *Facebook audience insights*. [Online]. Available: https://www.facebook.com/business/insights/tools/audience-insights, (accessed: 07.03.2022).

[10] Facebook, *About instagram insights*. [Online]. Available: https://www.facebook.com/business/help/441651653251838?id=419087378825961, (accessed: 07.03.2022).

[11] Twitter, *Twitter analytics*. [Online]. Available: https://business.twitter.com/en/advertising/analytics.html, (accessed: 07.03.2022).

[12] BuzzSumo. [Online]. Available: https://buzzsumo.com/, (accessed: 07.03.2022).

[13] Falcon.io. [Online]. Available: https://www.falcon.io/, (accessed: 07.03.2022).

[14] Hootsuite. [Online]. Available: https://www.hootsuite.com/, (accessed: 07.03.2022).

[15] Brandwatch. [Online]. Available: https://www.brandwatch.com/, (accessed: 07.03.2022).

[16] A. Bolioli, F. Salamino, and V. Porzionato, "Social media monitoring in real life with blogmeter platform.," *ESSEM@ AI* IA*, vol. 1096, pp. 156–163, 2013.

[17] L. Napalkova, P. Aragón, and J. C. C. Robles, "Big data-driven platform for cross-media monitoring," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 2018, pp. 392–399.

[18] J. Rogstadius, M. Vukovic, C. A. Teixeira, V. Kostakos, E. Karapanos, and J. A. Laredo, "Crisistracker: Crowdsourced social media curation for disaster awareness," *IBM Journal of Research and Development*, vol. 57, no. 5, pp. 4–1, 2013.

[19] D. Cameron, G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins, and R. Falck, "Predose: A semantic web platform for drug abuse epidemiology using social media," *Journal of biomedical informatics*, vol. 46, no. 6, pp. 985–997, 2013.

[20] . M. P. Gruzd A., *Communalytic: A research tool for studying online communities and online discourse*. [Online]. Available: https://communalytic.com/, (accessed: 14.04.2022).

[21] L. Meneses, "Netlytic," *Early Modern Digital Review*, vol. 2, no. 1, 2019.

[22] Texifter, *Discovertext*. [Online]. Available: https://discovertext.com/, (accessed: 14.03.2022).

[23] T. Elliott, *Socialbearing*. [Online]. Available: https://socialbearing.com/, (accessed: 14.03.2022).

[24] SocioViz. [Online]. Available: https://socioviz.net/SNA/eu/sna/login.jsp, (accessed: 14.03.2022).

[25] R. Ackland *et al.*, "Virtual observatory for the study of online networks (voson) software for collecting and analysing online networks (2004-2018)," 2018.

[26] D. Choi, Z. Matni, and C. Shah, "Socrates 2.0: Bridging the gap between researchers and social media data through natural language interactions," *Proceedings of the Association for Information Science and Technology*, vol. 52, no. 1, pp. 1–4, 2015.

[27] M. Thelwall, "Social web text analytics with mozdeh," *Mozdeh*, pp. 1–35, 2018.

[28] P. Brooker, J. Barnett, and T. Cribbin, "Doing social media analytics," *Big Data & Society*, vol. 3, no. 2, p. 2 053 951 716 658 060, 2016.

[29] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, and J. Conejero, "Cosmos: Towards an integrated and scalable service for analysing social media on demand," *International Journal of Parallel, Emergent and Distributed Systems*, vol. 30, no. 2, pp. 80–100, 2015.

[30] A. Guille, C. Favre, H. Hacid, and D. A. Zighed, “Sondy: An open source platform for social dynamics mining and analysis,” in *Proceedings of the 2013 ACM SIGMOD international conference on management of data*, 2013, pp. 1005–1008.

[31] C. Wang, L. Marini, C.-L. Chin, N. Vance, C. Donelson, P. Meunier, and J. T. Yun, “Social media intelligence and learning environment: An open source framework for social media data collection, analysis and curation,” in *2019 15th International Conference on eScience (eScience)*, IEEE, 2019, pp. 252–261.

[32] S. Srivastava and Y. N. Singh, “Big social media analytics: Applications and challenges,” in *Computer Networks, Big Data and IoT*, Springer, 2021, pp. 239–250.

[33] F. Emmert-Streib, O. P. Yli-Harja, and M. Dehmer, “Data analytics applications for streaming data from social media: What to predict?” *Frontiers in big Data*, p. 2, 2018.

[34] D. Camacho, M. V. Luzón, and E. Cambria, *New trends and applications in social media analytics*, 2021.

[35] W. He, H. Wu, G. Yan, V. Akula, and J. Shen, “A novel social media competitive analytics framework with sentiment benchmarks,” *Information & Management*, vol. 52, no. 7, pp. 801–812, 2015.

[36] E. Kalampokis, E. Tambouris, and K. Tarabanis, “Understanding the predictive power of social media,” *Internet Research*, 2013.

[37] C. M. Olszak, “An overview of information tools and technologies for competitive intelligence building: Theoretical approach,” *Issues in Informing Science and Information Technology*, vol. 11, no. 1, pp. 139–153, 2014.

[38] M. Asadi and A. Agah, “Characterizing user influence within twitter,” in *International Conference on P2P, Parallel, Grid, Cloud and Internet Computing*, Springer, 2017, pp. 122–132.

[39] Z. Z. Alp and Ş. G. Öğüdücü, “Influence factorization for identifying authorities in twitter,” *Knowledge-Based Systems*, vol. 163, pp. 944–954, 2019.

[40] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, “Measuring user influence in twitter: The million follower fallacy,” in *Proceedings of the international AAAI conference on web and social media*, vol. 4, 2010, pp. 10–17.

[41] F. Riquelme and P. González-Cantergiani, “Measuring user influence on twitter: A survey,” *Information processing & management*, vol. 52, no. 5, pp. 949–975, 2016.

[42] I. Anger and C. Kittl, “Measuring influence on twitter,” in *Proceedings of the 11th international conference on knowledge management and knowledge technologies*, 2011, pp. 1–4.

[43] J. Jacobson, A. Gruzd, and Á. Hernández-García, “Social media marketing: Who is watching the watchers?” *Journal of Retailing and Consumer Services*, vol. 53, p. 101 774, 2020.

[44] S. A. Mirlohi Falavarjani, J. Jovanovic, H. Fani, A. A. Ghorbani, Z. Noorian, and E. Bagheri, “On the causal relation between real world activities and emotional expressions of social media users,” *Journal of the Association for Information Science and Technology*, vol. 72, no. 6, pp. 723–743, 2021.

[45] L. Lefsrud, C. Westbury, J. Keith, and G. Hollis, “A basis for genuine dialogue: Developing a science-based understanding of public/industry communication,” *Phase I Report Prepared for the Alberta Chamber of Resources*, 2015.

[46] A. Malik, A. Johri, R. Handa, H. Karbasian, and H. Purohit, “How social media supports hashtag activism through multivocality: A case study of# ilooklikeanengineer,” *First Monday*, 2018.

[47] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, “Want to be retweeted? large scale analytics on factors impacting retweet in twitter network,” in *2010 IEEE second international conference on social computing*, IEEE, 2010, pp. 177–184.

[48] S. Stieglitz and L. Dang-Xuan, “Political communication and influence through microblogging–an empirical analysis of sentiment in twitter messages and retweet behavior,” in *2012 45th Hawaii international conference on system sciences*, IEEE, 2012, pp. 3500–3509.

[49] A. Kumar and G. Garg, “Sentiment analysis of multimodal twitter data,” *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24 103–24 119, 2019.

[50] A. Sapountzi and K. E. Psannis, “Social networking data analysis tools & challenges,” *Future Generation Computer Systems*, vol. 86, pp. 893–913, 2018.

[51] J. Bronstein, T. Gazit, O. Perez, J. Bar-Ilan, N. Aharony, and Y. Amichai-Hamburger, “An examination of the factors contributing to participation in online social platforms,” *Aslib Journal of Information Management*, 2016.

[52] S. Srivastava, M. K. Singh, and Y. N. Singh, “Social media analytics: Current trends and future prospects,” in *Communication and Intelligent Systems*, Springer, 2021, pp. 1005–1016.

[53] W. Y. Ayele and G. Juell-Skielse, “Social media analytics and internet of things: Survey,” in *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*, 2017, pp. 1–11.

[54] P. Melville, V. Sindhwani, and R Lawrence, “Social media analytics: Channeling the power of the blogosphere for marketing insight,” *Proc. of the WIN*, vol. 1, no. 1, pp. 1–5, 2009.

[55] S. B. Abkenar, M. H. Kashani, E. Mahdipour, and S. M. Jameii, “Big data analytics meets social media: A systematic review of techniques, open issues, and future directions,” *Telematics and Informatics*, vol. 57, p. 101 517, 2021.

[56] N. A. Ghani, S. Hamid, I. A. T. Hashem, and E. Ahmed, "Social media big data analytics: A survey," *Computers in Human Behavior*, vol. 101, pp. 417–428, 2019.

[57] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[58] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, "An overview of sentiment analysis in social media and its applications in disaster relief," *Sentiment analysis and ontology engineering*, pp. 313–340, 2016.

[59] J. Scott, "Social network analysis: Developments, advances, and prospects," *Social network analysis and mining*, vol. 1, no. 1, pp. 21–26, 2011.

[60] S. C. Lewis, R. Zamith, and A. Hermida, "Content analysis in an era of big data: A hybrid approach to computational and manual methods," *Journal of broadcasting & electronic media*, vol. 57, no. 1, pp. 34–52, 2013.

[61] M. Rooduijn and T. Pauwels, "Measuring populism: Comparing two methods of content analysis," *West European Politics*, vol. 34, no. 6, pp. 1272–1283, 2011.

[62] N. Clarke, P. Foltz, and P. Garrard, "How to do things with (thousands of) words: Computational approaches to discourse analysis in alzheimer's disease," *Cortex*, vol. 129, pp. 446–463, 2020.

[63] M. Huisman and M. A. van Duijn, "Software for statistical analysis of social networks," in *The Sixth International Conference on Logic and Methodology*, Amsterdam The Netherlands, 2004.

[64] W. Fan and M. D. Gordon, "The power of social media analytics," *Communications of the ACM*, vol. 57, no. 6, pp. 74–81, 2014.

[65] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics–challenges in topic discovery, data collection, and data preparation," *International journal of information management*, vol. 39, pp. 156–168, 2018.

[66] S. Stieglitz and L. Dang-Xuan, "Social media and political communication: A social media analytics framework," *Social network analysis and mining*, vol. 3, no. 4, pp. 1277–1291, 2013.

[67] R. Sellami, F. Zalila, A. Nuttinck, S. Dupont, J.-C. Deprez, and S. Mouton, "Fadi-a deployment framework for big data management and analytics," in *2020 IEEE 29th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, IEEE, 2020, pp. 153–158.

[68] R. K. Singh and H. K. Verma, "Redis-based messaging queue and cache-enabled parallel processing social media analytics framework," *The Computer Journal*, 2020.

[69] T. A. S. Foundation, *Apache kafka.* [Online]. Available: https://kafka.apache.org/, (accessed: 14.04.2022).

[70] T. A. S. Foundation, *Apache flume*. [Online]. Available: https://flume.apache.org/, (accessed: 14.04.2022).

[71] T. A. S. Foundation, *Apache nifi*. [Online]. Available: https://nifi.apache.org/, (accessed: 14.04.2022).

[72] T. A. S. Foundation, *Apache pulsar*. [Online]. Available: https://pulsar.apache.org/, (accessed: 14.04.2022).

[73] T. A. S. Foundation, *Spark structured streaming*. [Online]. Available: https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html, (accessed: 14.04.2022).

[74] T. A. S. Foundation, *Apache storm*. [Online]. Available: https://storm.apache.org/, (accessed: 14.04.2022).

[75] T. A. S. Foundation, *Apache samza*. [Online]. Available: https://samza.apache.org/, (accessed: 14.04.2022).

[76] Google, *Google dataflow*. [Online]. Available: https://cloud.google.com/dataflow, (accessed: 14.04.2022).

[77] Amazon, *Amazon kinesis*. [Online]. Available: https://aws.amazon.com/kinesis/, (accessed: 14.04.2022).

[78] T. A. S. Foundation, *Apache spark*. [Online]. Available: https://spark.apache.org/, (accessed: 14.04.2022).

[79] T. A. S. Foundation, *Https://flink.apache.org/*. [Online]. Available: https://flink.apache.org/, (accessed: 14.04.2022).

[80] Crate.io, *Cratedb*. [Online]. Available: https://crate.io/, (accessed: 14.04.2022).

[81] Timescale, *Timescaledb*. [Online]. Available: https://www.timescale.com/, (accessed: 14.04.2022).

[82] Microsoft, *Azure cosmos db*. [Online]. Available: https://azure.microsoft.com/en-us/services/cosmos-db/, (accessed: 14.04.2022).

[83] T. L. Foundation, *Prometheus*. [Online]. Available: https://prometheus.io/, (accessed: 14.04.2022).

[84] C. Labs, *Cockroachdb*. [Online]. Available: https://www.cockroachlabs.com/product/, (accessed: 14.04.2022).

[85] InfluxData, *Influxdb*. [Online]. Available: https://www.influxdata.com/get-influxdb/, (accessed: 14.04.2022).

[86] T. A. S. Foundation, *Apache pinot*. [Online]. Available: https://pinot.apache.org/, (accessed: 14.04.2022).

[87] T. A. S. Foundation, *Apache superset*. [Online]. Available: https://superset.apache.org/, (accessed: 14.04.2022).

[88] Slaesforce, *Tableau*. [Online]. Available: https://trust.tableau.com/, (accessed: 14.04.2022).

[89] *Redash.* [Online]. Available: https://redash.io/, (accessed: 14.04.2022).

[90] *Metabase.* [Online]. Available: https://www.metabase.com/, (accessed: 14.04.2022).

[91] ZingSoft, *Zingchart.* [Online]. Available: https://www.zingchart.com/, (accessed: 14.04.2022).

[92] U. I. D. Lab, *Vega.* [Online]. Available: https://vega.github.io/vega/, (accessed: 14.04.2022).

[93] M. Bostock, *D3.* [Online]. Available: https://d3js.org/, (accessed: 14.04.2022).

[94] Chart.js, *Chart.js.* [Online]. Available: https://www.chartjs.org/, (accessed: 14.04.2022).

[95] Google, *Google charts.* [Online]. Available: https://developers.google.com/chart, (accessed: 14.04.2022).

[96] V. Lab, *Vosonsml.* [Online]. Available: https://vosonlab.github.io/vosonSML/, (accessed: 14.04.2022).

[97] P. A. R. Group, *Socrates.* [Online]. Available: https://socrates.peopleanalytics.org/, (accessed: 14.04.2022).

[98] A. Gruzd, "Netlytic: Software for automated text and social network analysis," *Diakses dari http://netlytic. org*, 2016.

[99] S. C. R. Group, *Mozdeh.* [Online]. Available: http://mozdeh.wlv.ac.uk/index.html, (accessed: 14.04.2022).

[100] A. N. Liaropoulos, "The challenge of social media for the intelligence community," *Journal of Mediterranean and Balkan intelligence*, vol. 1, no. 1, pp. 5–14, 2013.

[101] B. Senekal and E. Kotzé, "Open source intelligence (osint) for conflict monitoring in contemporary south africa: Challenges and opportunities in a big data context," *African Security Review*, vol. 28, no. 1, pp. 19–37, 2019.

[102] D. Omand, C. Miller, and J. Bartlett, "Towards the discipline of social media intelligence," in *Open source intelligence in the twenty-first century*, Springer, 2014, pp. 24–43.

[103] C. Hobbs, M. Moran, and D. Salisbury, *Open source intelligence in the twenty-first century: new approaches and opportunities.* Springer, 2014.

[104] N. Antonius and L Rich, "Discovering collection and analysis techniques for social media to improve public safety," *The international technology management review*, vol. 3, no. 1, pp. 42–53, 2013.

[105] R. Li, J. Crowe, D. Leifer, L. Zou, and J. Schoof, "Beyond big data: Social media challenges and opportunities for understanding social perception of energy," *Energy Research & Social Science*, vol. 56, p. 101 217, 2019.

[106] R. RAICU *et al.*, "The emergence of social media intelligence," *Romanian Intelligence Studies Review*, no. 14, pp. 181–196, 2015.

[107] K. L. O'Halloran, G. Pal, and M. Jin, "Multimodal approach to analysing big social and news media data," *Discourse, Context & Media*, vol. 40, p. 100 467, 2021.

[108] V. Ponizovskiy, M. Ardag, L. Grigoryan, R. Boyd, H. Dobewall, and P. Holtz, "Development and validation of the personal values dictionary: A theory–driven tool for investigating references to basic human values in text," *European Journal of Personality*, vol. 34, no. 5, pp. 885–902, 2020.

[109] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 168–177.

[110] T. Engelthaler and T. T. Hills, "Humor norms for 4,997 english words," *Behavior research methods*, vol. 50, no. 3, pp. 1116–1124, 2018.

[111] C. Westbury and G. Hollis, "Wriggly, squiffy, lummox, and boobs: What makes some words funny?" *Journal of Experimental Psychology: General*, vol. 148, no. 1, p. 97, 2019.

[112] E. Nikitin, *Redditscore*, https://github.com/crazyfrogspb/RedditScore, 2018.

[113] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[114] V. Singh, "Replace or Retrieve Keywords In Documents at Scale," *ArXiv e-prints*, Oct. 2017. arXiv: 1711.00046 [cs.DS].

[115] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the international AAAI conference on web and social media*, vol. 8, 2014, pp. 216–225.

[116] T. Binder, *How to calculate twitter impressions and reach.* [Online]. Available: https://www.tweetbinder.com/blog/twitter-impressions/, (accessed: 14.04.2022).

[117] A. Guilds, *7 best ways to calculate engagement rate on social media.* [Online]. Available: https://www.authorsguilds.com/how-to-calculate-social-media-engagement-rate/, (accessed: 14.04.2022).

[118] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, and M. Tesconi, "Cross-media learning for image sentiment analysis in the wild," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 308–317. DOI: 10.1109/ICCVW.2017.45.

[119] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2020, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802. [Online]. Available: https://doi.org/10.1109/ASYU50717.2020.9259802.

[120] L. Yin, *Smappnyu/urlexpander: Initial release*, Aug. 2018. DOI: 10.5281/zenodo.1345144. [Online]. Available: https://doi.org/10.5281/zenodo.1345144.

[121] T. A. S. Foundation, *Apache zookeeper*. [Online]. Available: https://zookeeper.apache.org/, (accessed: 14.04.2022).

[122] H. Liu, "Rapid application configuration in amazon cloud using configurable virtual appliances," in *Proceedings of the 2011 ACM Symposium on Applied Computing*, 2011, pp. 147–154.

[123] C. A. G. Gutierrez, A. Whittaker, K. M. Patenio, J. Gehman, L. M. Lefsrud, D. Barbosa, and E. Stroulia, "Analyzing and visualizing twitter conversations," in *Proceedings of the 31st Annual International Conference on Computer Science and Software Engineering*, 2021, pp. 4–13.

Appendix A: SoDa-TAP APIs

The set of APIs available in *SoDa-TAP* as seen in Table A.1, requires for different parameters to return specific data results. They return ”Success” or ”Error” from any of the endpoints, depending on the result from a call. A valid call will return a result as a JSON, since is the standard way to communicate between server and client. When there is an invalid call, the API returns an error number and a brief error message. Both conditions will be able to happen if the case is that the service is running.

Table A.1: SoDa-TAP API Endpoints.

Endpoint	**Description**	**Parameters**	**Response**
FastAPI			
/select/	Select data in a paginated fashion.	[table, id_column_name, time_column_name, columns, lastID, lastTime]	Success or, Error
/select_offset/	Get pieces of data by offset rows.	[table, columns, size, offset]	Success or, Error
/select_ids/	Get data by IDs.	[table, columns, id_column_name, ids]	Success or, Error
/select_date_range/	Get data by date range.	[table, columns, time_column_name, start, end]	Success or, Error
Socket.IO			
crate_query	A SQL query to execute in the database.	[socket.ioID, query]	Success or, Error