Analyzing and Visualizing Twitter Conversations

Candelario A. Gutierrez Gutierrez† (cagutier@ualberta.ca), Andrea Whittaker † (amwhitta@ualberta.ca), Katherine Mae Patenio † (patenio@ualberta.ca), Joel Gehman * (jgehman@gwu.edu), Lianne M. Lefsrud †† (lefsrud@ualberta.ca), Denilson Barbosa † (denilson@ualberta.ca), and Eleni Stroulia† (stroulia@ualberta.ca)

† Computing Science, University of Alberta, Edmonton Alberta Canada
* Strategic Management and Public Policy; George Washington University, Washington D.C. United States
†† Chemical and Materials Engineering, University of Alberta, Edmonton Alberta Canada

ABSTRACT

Social media platforms are public venues where conversations about issues of public interest take place. Much recent research has been devoted to evaluating the degree to which online conversations capture public opinion on issues of broad societal interest. We describe a robust and scalable platform to support such studies. Our platform allows the analysis of three semantic aspects of tweets, namely the personal values, sentiment, and humor expressed in them, as well as the public's engagement with them. In addition, it aggregates these indicators at the level of tweet authors to shed light on the activities and style of influencers of public opinion. Finally, it offers rich visualizations to enable users to gain insights on their datasets. We demonstrate the usefulness of our platform with two case studies: (a) analyzing the fragmented narratives around established (hydro, oil and gas, coal, nuclear) and new (solar, wind, geothermal, biomass) energy sources; and (b) comparing the social-media brands of academic institutions.

CCS CONCEPTS

semantic analysis, sentiment, data visualization, case studies

KEYWORDS

social media, personal values, humor, energy narratives, academic institutions' brands

1 Introduction

Recently, conversations about public interest issues have shifted. First, they have shifted in location: from mediated traditional media sources to unmediated online media, blogs, and webpages. For many controversial issues such as pipeline developments, or climate change, or COVID-19 public-safety measures, rather than a central organization defining its stakeholders, individuals and organizations are self-identifying with issues, opting into increasingly unmediated debate, and actively evaluating risks and rejecting the positions of organizations and even regulators [23, 24]. Second, online conversations — meaning, value, and emotions — have lasting financial consequences [22]: a tweet from a disgruntled customer about a corporation's poor service can cause a public-opinion fallout and push a multitude of clients away from the brand. Third, online stakeholder activism can create an existential crisis for organizations that use purely technical methods of delineating their context and identifying and evaluating their risks. Technical approaches to risk evaluation, like cost-benefit analysis "generally cannot resolve strong differences in value judgments that are often present in controversial projects" [28]. Finally, online conversations about one topic are inter-related with many conversations on other related topics. For example, discussions about invest/divest intermingle with policy discussions about climate change [32].

These shifts to online discussion and mediation affect a broad set of outcomes, including patterns of collaboration, strategy, and behaviour, resulting in substantial interest in developing computational methods for understanding the conversations that take place on social platforms [41]. These include algorithms for analyzing the structure of the social network, i.e., its key influencers, the network embedded communities, and their evolution [3, 17]; the topics of the conversations and their spatiotemporal trends [33]; and the emotional valence and arousal of the various contributions [29].

As the research matures, more studies are being conducted on social platforms, and especially on Twitter, as it offers fairly open access to its data for which numerous software libraries are available on code-sharing platforms. Indicatively, a cursory search on Google Scholar at the time of writing with variations on the search phrase "twitter analysis for covid" returns more than 200 publications in 2021 alone. Evidently a platform to conduct such studies would enable a more efficient exchange of data and findings, and the replicability of the related research.

To that end, we have developed a robust and scalable platform that analyzes three semantic aspects of tweets: personal values, sentiment, and humor expressed, and public engagement with them. From this, we aggregate these indicators to the level of Twitter users to examine their overall "style" of contributions and their relative influence. Finally, we develop complementary rich visualizations to expose insights on such datasets. We demonstrate the usefulness of our platform with two case studies: (a) analyzing the fragmented narratives around established (hydro, oil and gas, coal, nuclear) and new (solar, wind, geothermal, biomass) energy sources; and (b) comparing the social-media brands of academic institutions.

The rest of this paper is organized as follows. Section 2 briefly summarizes the basic techniques integrated in our platform. Section 3 describes its software architecture, and the visualizations currently implemented in our platform and the questions they are meant to help answer. Section 4 describes our experience with two different data sets. Finally, Section 5 concludes with a summary of our work and our future plans.

2 Background and Related Research

2.1 Dictionary-based Analysis

A key methodology for semantic analysis of text is the use of standardized dictionaries, expertly curated to associate words with interesting semantic dimensions. Dictionary-based text analysis is conceptually simple and efficient. Once a dictionary has been curated and demonstrated to be valid, software systems can easily incorporate it in their processing, in a manner that can be scalable to large text corpora, through data parallelism.

Our platform incorporates three dictionaries. The personal values dictionary created by [37] can be seen as a model of the personal values expressed in text but compared to ours it does not take context of the words into consideration. Values have been a topic of longstanding and continuing interest across a variety of social scientific disciplines [8, 11, 13, 35, 40, 42, 45, 50]. Within this stream, a key contribution has been Schwartz's [42, 43, 44] work, which focuses on value priorities, theorizing their role in influencing behavioral orientations and choices such as ideologies, attitudes, and actions of individuals. Within this perspective, values are considered essential to self-understanding as well as criteria used to select and justify action and to evaluate actors and events [40, 42]. Values also are meaningful and relevant at both the individual and collective levels [18, 19]. Values also play a key role in shaping social reality and structures [21, 50]. Finally, although many scholars define values as conceptions of the desirable, others define them in terms of what is undesirable, e.g. [10]. In particular, our model builds on Schwartz's circumplex model which nests 57 discrete values (e.g., tolerance, dominance, dependability) within 10 aggregate dimensions (e.g., self-protection versus growth, personal focus versus social focus).

A humor dictionary has been created by Westbury and Hollis [51], based on predictions of the funniness of words. They analyzed the semantic, phonological, orthographic, and frequency factors that play a role in the judgments of humor. From this, they were able to predict the original humor rating norms and ratings for previously unrated words with greater reliability. Their findings are consistent with several theories of humor, especially the incongruity theory, which suggests that individuals' experience of humor is proportional to the degree to which expectations are violated. Humor increases the likelihood of persuasion, knowledge, and attitudes [49].

The sentiment dictionary of Hu and Liu [15], composed of around 6,800 words, is categorized by positive and negative emotional valence. It was first created from opinion based online content and we found it useful for our methodology because it was designed with social media applications in mind.

2.2 Aspect-Based Sentiment Analysis

Aspect-based sentiment analysis aims at identifying the aspects of the entities mentioned in text and determine the sentiment of the author toward those aspects. A popular approach to this task is to use dependency parsing as a way of isolating the aspect of interest along with its modifying words [34, 52]. The emotional valence of the modifying words is then evaluated through lexicons or rule-based models. This is the approach to that we take in our work, as alternative models are highly domain specific and require large amounts of labelled data.

2.3 Social-Platforms Analysis

Social media data have been widely used to support different domains of studies such as marketing and consumer behavior [12], smart cities [26], disaster management [20], health [4] and more. Hence, different frameworks and platforms have been proposed and have shown to be favorable for social media analysis [1, 2, 7, 30, 36, 38, 47]. However, they all work with complex models that require high volume of labelled data and other complexities that make studies hard to replicate in another domain. Furthermore, several analysis results are presented through a specific package for a programming language, making it hard for the inexperienced audience to interact with the insights that are not fully understood because of visualizations' incorrect use. Finally, most studies do not fully exploit the potential of conversational data through the measurement of engagement and the metadata that can be extracted from it.

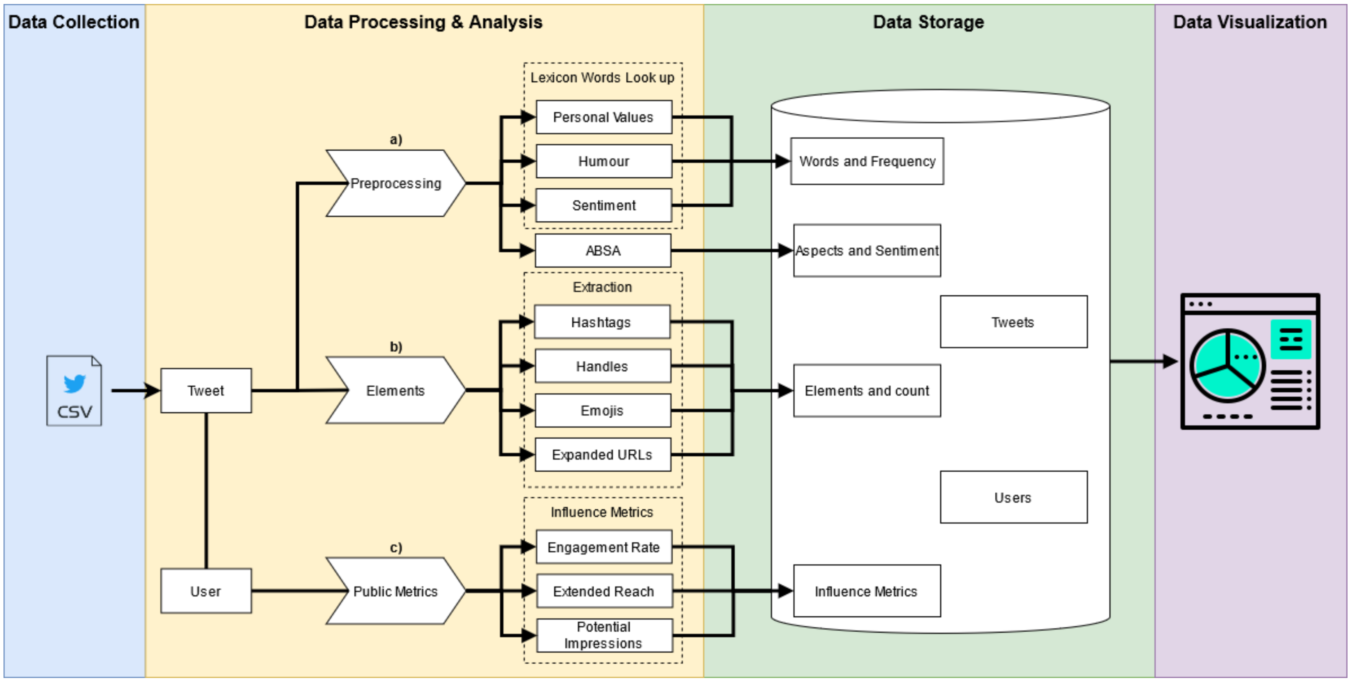


Figure 1. Architecture of the Platform

On the other hand, there are several studies demonstrating that lexicon-based methodologies can perform well and reduce complexity to understand semantics of text. McCaig et al. [27] analyzed Reddit conversations to measure the relevance of 4 concepts: fitness trackers, eating, body and exercise. Each concept was represented as a list of previously studied words. These word lists were then matched in the posted comments to obtain word frequencies per list. They also calculated other metrics such as the number of threads, the number of comments, the number of unique commenters and the average number of comments by each unique commenter.

Alomari and Mehmood [6] extracted a set of tweets in Arabic language coming from popular accounts that mention the same city's location and constructed 4 different dictionaries containing: street names of the city's location, synonyms of traffic, traffic reasons, and transportation words. They then used these dictionaries to find matches in the tweets that were used to create a tag per match. The resulting tags got presented in a cone chart to visualize volume, a pie chart to show streets congestion, a horizontal bar chart to show posted tweets per hour, and a cloud chart to visualize the traffic synonyms, reasons, and transportation words.

Lingiardi et al. [25] extracted data from Facebook and Twitter to locate zones that hate speech was most widely happening around Italy. They achieved this by gathering a set of previously studied insulting words in Italian, used against different minority groups. This set of words would later be used against the data extracted from the social platforms to find their corresponding match, and in case there were ambiguous words, they would apply a semantic tagger and a sentiment analyzer to conserve only hate words. A geolocation filter was also applied to set boundaries while extracting the data. Their results were presented in a geographic map where they denoted the degrees of hate speech through colors: green as none, yellow as moderate and red as frequent.

Al-Daihani and Abrahams [5] extracted Facebook data related to libraries from high performing schools present in Canada, the United Kingdom, Australia, and the United States. Their metrics of interest were the number of page likes, post likes and post comments. They defined engagement as a multiplication of likes and comments. Statistical analysis such as min, max, average, std. deviation and median were analyzed over the metrics of interest. They compared cumulative data per country where they got to compare metrics and they used unigrams and bigrams to calculate word frequency. All the results were showed in horizontal bar and line charts.

We propose a configurable platform alongside a methodology to work with social media data. Through the unification of lexicon-based and engagement analysis that can be interpreted through a useful set of visualizations.

3 Software Architecture

Social media is a prime example of semi-structured data consisting primarily of free-form texts, called posts, containing references to users (e.g., "at mentions" in Twitter), topics of interest (e.g., hashtags), or other posts (e.g., "re-tweets" in Twitter). Because each post has a unique identifier, social media corpora are best handled by modern key-value stores which are part of the no-SQL family. On the other hand, one also needs the help of a powerful query language to gather insights from aggregated data. Finally, social media corpora can easily grow to terabytes, implying that a distributed and easily parallelizable architecture is needed. For these reasons we chose CrateDB[1], a scalable key-value store, a tried and true highly performant open-source DBMS, as our back-end. Next, we explain how we handle the posts, starting from processing the texts for subsequent analysis.

Figure 1 diagrammatically depicts an overview of our platform architecture. A detailed description of how this platform works is presented below.

3.1 Data Collection

The data-collection process is given as input a period during which the tweets to be collected have appeared; a list of hashtags that should be included in the tweets, or a list of keywords that should be mentioned in them; and, optionally, a list of users whose tweets should be collected. In principle there are two ways to collect a set of tweets: (a) using Twitter's API full-archive search endpoint[2] or (b) using Twint[3] a scraper library that gathers tweets from users' timelines. The former is more reliable if the period is of importance, the latter when a list of users are of interest. Our platform includes two different Twitter clients to support both the above data-collection methods that export the collected tweets into a single CSV.

3.2 Data Processing & Analysis

This CSV is subsequently loaded into Apache Spark[4] in local mode inside a Docker[5] container. The platform implements three key processes.

a) A set of preprocessing pipelines transform the input tweets in "cleaner" versions appropriate for two types of subsequent analysis: lexicon-based and aspect-based. The first pipeline is used to support word look up based on three standard lexicons corresponding to a set of words converted into their base form with Stanza [39]. To normalize the tweets, the tweet-preprocessor[6] and gensim[7] libraries are used to remove stop words, handles, URLs, emojis, numbers, hashtags, punctuation and non-unicode characters, and all text is transformed to lowercase. As a final step, Spark NLP's[8] lemmatization module is used to convert all the tweets' words into their base form. Once tweets are normalized, they are processed to extract word frequencies depending on the dictionary.

Another pipeline prepares the tweets for the aspect-based sentiment analysis by removing numbers, URLs and twitter handles (as they do not convey opinions or sentiment and can thus be ignored); lowercasing all words, except those with all letters capitalized (as sentiment analysis tools see that as a signal of strong arousal of emotion); and splitting hashtags into words using the CrazyTokenizer [31] (as unlike URLs and handles, hashtags can express sentiment); and preserving punctuation (as sentiment analysis tools often rely on them as features).

The aspect-based sentiment analysis process starts with dependency parsing, using SpaCy [14], to reveal the relationships between words that modify the meaning of other words. More precisely, the modifiers of each aspect found in the tweet are examined in search of emotive words that directly relate to the aspect. When multiple aspects are found in the tweet, they are separated in the dependency structure and their specific modifiers can later be analyzed for sentiment independently. When no modifiers are attached to an aspect, the emotive words within a window of 3 words preceding and 3 words following the aspect are collected. We use VADER [16] to measure sentiment, as it is rule-based and has been shown highly effective for social media. We use the compound polarity score from VADER to quantify the sentiment of the noun chunk, using traditional ranges of (-1.0,-0.5), (-0.5,0.5), (0.5,1) to mean negative, neutral, and positive respectfully.

b) A set of extraction functions collect and quantify elements present in the text such as hashtags, handles, emojis and expanded URLs.

c) Engagement analysis is finally performed to quantify the impact of each tweet using three metrics. The *engagement rate* considers 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. The *extended reach* metric is based on the number of retweets divided by the total number of tweets by the author, multiplied by 100. The *user's potential impressions* metric is based on the number of user's followers multiplied by the total number of the user's tweets count.

The results from all the processes before they reach the database are stored in a separate Spark SQL[9] DataFrame based on the relationship shown in Figure 2. The tweets DataFrame columns hold (i) the tweet's text; (ii) its public

[1] https://crate.io/

[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://spark.apache.org/

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

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

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

[8] https://www.johnsnowlabs.com/spark-nlp/

[9] https://spark.apache.org/sql/

metrics: retweet_count, reply_count, like_count, quote_count; (iii) created_at timestamp; (iv) words and frequency of the lexicons look up, (v) aspects found and their sentiment score; (vi) tweet's sentiment; (vii) elements extracted and their count; (viii) influence metrics; and (ix) a reference id of the user's table. The users DataFrame columns hold (i) the user's id; (ii) its public metrics: followers_count, following_count, tweet_count, listed_count; (iii) verified boolean; (iv) bio and (v) screen_name.

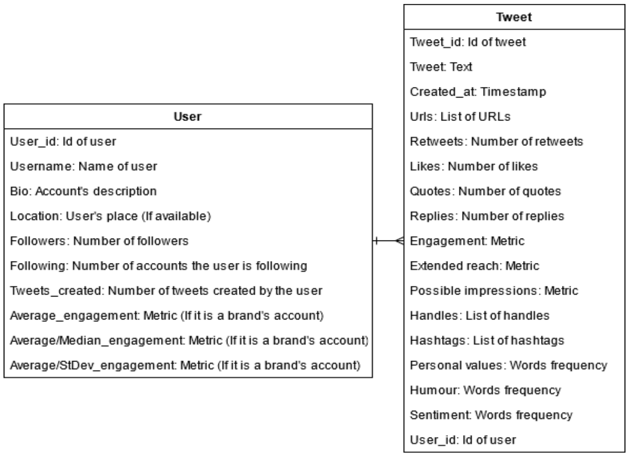

User
User_id: Id of user
Username: Name of user
Bio: Account's description
Location: User's place (If available)
Followers: Number of followers
Following: Number of accounts the user is following
Tweets_created: Number of tweets created by the user
Average_engagement: Metric (If it is a brand's account)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

3.3 Data Storage

CrateDB is used to store the generated DataFrames of tweets and users. It is a horizontally scalable time-series database for data consultation and time filtering. Some characteristics that make it a good fit are that it is schema-flexible, can hold multiple columns, exposes a Representational State Transfer (REST) API and a PostgreSQL wire protocol. The last two make it compatible with a broad variety of external applications and use cases and since Spark can work with Java Database Connectivity (JDBC) drivers, the connection between Spark and CrateDB can be stablished in two ways, with the PostgreSQL JDBC[10] driver or through a connection with SQLAlchemy[11] using Pandas[12].

3.4 Data Visualization

Two groups of visualizations are available to contextualize (a) the frequency of personal values, sentiment, and humor in tweets, and (b) the mention of terms associated with an aspect and what combinations of them are popular in tweets. They are all configured and built with ZingChart[13], served through the web and can communicate directly with the database thanks to the compatibility of both ends to communicate through API calls.

In both groups, the visualizations are rich in description and more detailed information is available while interacting with them in a web browser.

3.4.1 Personal Values, Sentiment, and Humor

The first group of visualizations consist of a treemap, line and area, and a streamgraph chart. The purpose of the treemap is to visualize the volume of word frequency associated with a personal value based on different filtering options: (a) cumulative of all the dataset; (b) day, month, or year range comparison and (c) specific dates range comparison. For all filtering options there is a box per personal value with an assigned color. Inside each box are internal boxes that represent the words of each personal value. Depending on the box, there is a text label that denotes a personal value or a word, with the size of the box representing the volume of the frequency count. If any of the filtering options are selected, the background of each word box will become red or green. Green if the end date has a higher frequency for a word compared to the start date, and red otherwise. If there is no difference, the color will be that of the personal value.

The line and area charts are meant to show cumulative frequency of words over time, per dictionary. In these charts the user can drag and select specific data points in time to zoom in or out, to see their distribution.

The streamgraph's functionality is similar to the line and area charts with the difference being that it is easier to spot their distribution.

3.4.2 Aspects and Terms

The second group of visualizations is composed of four different charts that focus on tweets that have aspects. Aspects are defined here as the attributes, features, components, or relevant considerations of energy sources and the energy industry in Canada. The aspect list was created by consulting Natural Resource Canada's *Energy Factbook*, industry experts' listing of keywords and projects, and Wikipedia entries. Some of these aspects are specific to a few energy sources, while others are applicable to all sources. Once we determined our list of aspects, we searched ConceptNet [46] for related terms to improve the recall of tweets mentioning these aspects.

A pie chart shows the representative quantity of tweets with and without aspects. It is possible for a tweet to have more than one aspect term, whether those terms are associated with the same aspect or originate from different aspects. For instance, a tweet may contain the terms "protect" and "environmental"; "protect" is a term under the "safety" aspect, whereas "environmental" is a term under the "sustainability" aspect. If there is a tweet that has two terms from the same

[10] https://jdbc.postgresql.org/
[11] https://www.sqlalchemy.org/
[12] https://pandas.pydata.org/
[13] https://www.zingchart.com/

parent, such as “protect” and “security” from the “safety” aspect, this would be considered a combination, that is to say, a combination “protect, safety”.

A sunburst chart compares the number of tweets mentioning terms tied to an aspect, where each slice corresponds to its frequency. The inner ring represents all aspects and the outer one represents all combinations of terms for each aspect.

A treemap compares the number of tweets mentioning terms associated with an aspect. It shows the sentiment of not only each aspect, but also each combination of terms via the colour of the box. Colours include: (a) red for negative; (b) light red for slightly negative; (c) grey for neutral; (d) light green for slightly positive; (e) green for positive. We distinguish between slightly positive or negative sentiments with VADER’s range. Sentiments for term combinations are calculated by finding the mean compound of each term. Each mean compound found is then used to determine the overall mean compound of an aspect. The treemap can be filtered by (a) allowing the user to show or hide individual aspects on the treemap and (b) selecting “all” or “range” to explore how the frequency of tweets and term combinations have changed over time.

Finally, a violin chart compares the distribution and density of mean compounds which were used to calculate the sentiments of aspects and combinations of terms. Each violin represents an aspect while mean compounds are displayed on the y-axis. This chart only displays all mean compounds found overall.

3.4.3 Tweets and Users Filtering Table

With FancyGrid[14], we created a tweets and users’ table. These tables have the same functionality with the only difference being the data represented in each of them. The user is able to (a) search the whole table and filter by keyword; (b) show/hide columns; (c) export the table data to a CSV; (d) filter per individual columns and in case of a numeric column, use the operators: <, >, <=, >=, != to compare and (e) see the median, average, and standard deviation of engagement of the tweets table. In these tables, tweets, users, and links can be clicked. There is a page number filter and a selector to pick how many rows to show. As an extra metric, there is a legend denoting the page number and the total number of records available.

4 Case Studies

To understand the potential of the platform we have conducted two different analyses. One that aims to understand conversations around a specific topic and another one that focuses on the exploration of metrics around a brand.

4.1 The Energy East Case Study

In the first use case we chose to explore what people had to say about the energy east pipeline. The project was announced on August 1, 2013 and became heavily debated among multiple groups on economic, political, environmental, and moral grounds. To be able to understand such perspectives and conversations we mainly looked for personal values and custom-defined aspects to look for sentiment, i.e., concepts that would let us understand how people were feeling, how they were expressing themselves online about it and how this evolved through time. In addition, we looked for humor and sentiment words.

We started with a Twitter query search based on the “#energyeast” hashtag between the dates of March 21, 2006 to June 17, 2021. It gave us 28,693 tweets, 111,091 retweets, 3,753 quotes and 4,780 replies with a total of 148,317 tweets, in the period from June 7, 2013 to June 16, 2021. After applying our preprocessing functions, lexicon dictionaries look up and engagement calculations, we filtered the resulting data to consider only tweets, quotes and replies for further analysis.

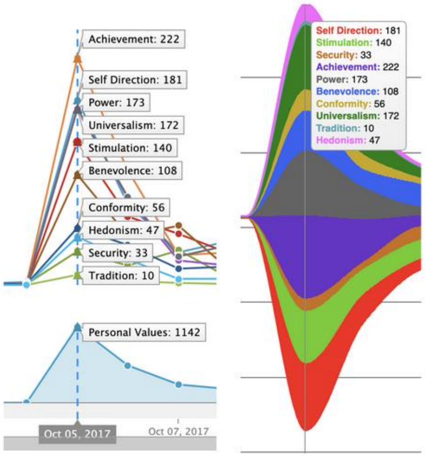


Figure 3. Line and Area, and Streamgraph Chart for Personal Values Dictionary

Using the line and area chart we were able to perceive three activity spikes that happened on January 22, 2016, January 27, 2016, and October 5, 2017, the last one being the most distinctive, as Figure 3 depicts. This date would then be our main point of focus for further analysis.

We decided to pick the top three personal values identified: achievement, self-direction, and power. Based on this set of words, we decided to investigate further with the personal values treemap as seen in Figure 4, filtered to only this specific date. From this visualization we noticed the words: job (achievement - 34 words); decision (self-direction - 50 words) and victory (power - 33 words).

[14] https://fancygrid.com/

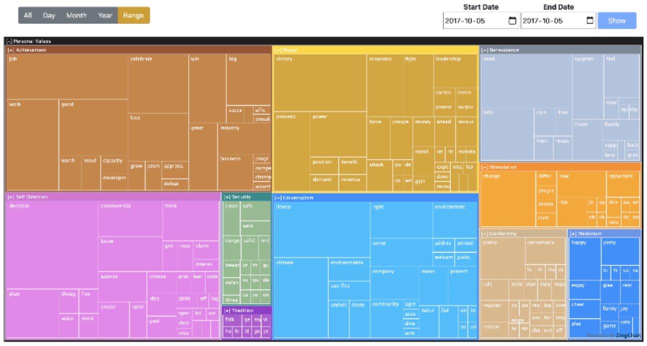


Figure 4. Personal Values Treemap

To further understand the context, we went back to the line and area, and the streamgraph charts but now applied to the sentiment dictionary as seen in Figure 5. This showed us that there were more negative (53.2%) than positive (46.8%) words.

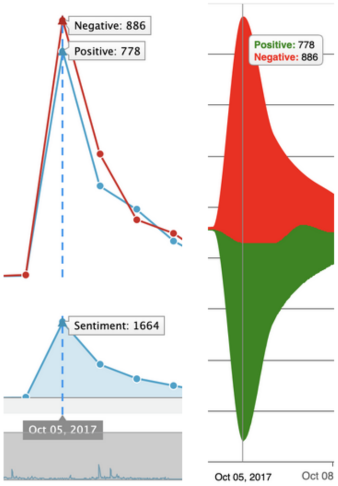
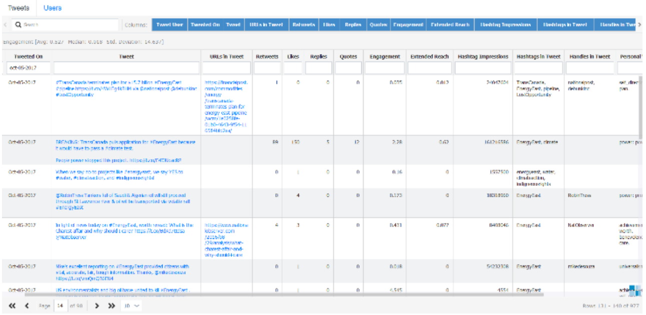


Figure 5. Line and Area, and Streamgraph Chart for the Sentiment Dictionary

More analysis granularity was achieved with the help of the sunburst and treemap charts for aspects. The results that we got from filtering both charts with the same date are shown in Figures 7 and 8. They showed us that the top 3 aspects found from the conversations were about: employment (29.7% - jobs with 32 tweets); cost (18.8% - worth with 6 tweets) and safety (14.1% - safe and protect with 4 tweets). As it is important to know detailed information about the aspects in a specific date, it is also helpful to visualize distribution from the overall data. We accomplished this with the pie and the violin charts shown in Figure 9 and 10 respectively. For the first one, out of 37,226 tweets (tweets, replies and quotes), only 4,793 tweets (12.9%) have aspect sentiments. Each tweet within the group of 4,793 contains at least one aspect term. For the second one, it showed us that there is high variability of aspects sentiment that refer to safety, sustainability, and employment.

One more analysis that we were able to perform was to get how many users participated in the conversation. This was achieved through the users table that looks like Figure 6. From this table we were able to spot that a total of 977 tweets were done in this same date: 695 tweets; 169 quotes and 113 replies. Where 629 unique users participated.



engagement [Avg: 0.527 Median: 0.018 Std. Deviation: 14.637]

Tweeted On	Tweet	URLs in Tweet	Retweets	Likes	Replies	Quotes	Engagement	Extended Reach	Hashtag Impressions	Hashtags in Tweet	Handles in Tweet	Personal
oct-05-2017												
Oct-05-2017	#TransCanada terminates plan for a $15.7 billion #EnergyEast #pipeline [illegible] via @nationalpost @debunking #LostOpportunity	https://financialpost.com/commodities/energy/transcanada-terminates-plan-for-energy-east-pipeline/wcm/[illegible]	1	0	0	0	0.015	0.012	24047604	TransCanada, EnergyEast, pipeline, LostOpportunity	nationalpost, debunking	self_direc… pan…
Oct-05-2017	BREAKING: TransCanada pulls application for #EnergyEast because it would have to pass a #climate test. People power stopped this project. [illegible]		89	150	5	12	2.28	0.62	161216586	EnergyEast, climate		power; po…
Oct-05-2017	When we say no to projects like #energyeast, we say YES to #water, #climateaction, and #indigenousrights		0	1	0	0	0.16	0	1557500	energyeast, water, climateaction, indigenousrights		
Oct-05-2017	@RobinThree Tankers full of Saudi & Algerian oil which proceed through St Lawrence river & oil will be transported via [illegible] #energyeast		0	4	0	0	0.173	0	10318000	EnergyEast	RobinThree	power; pro…
Oct-05-2017	In light of news today on #EnergyEast, worth a read: What is the charest affair and why should I care? [illegible] @NatObserver	https://www.nationalobserver.com/2016/08/29/analysis/what-charest-affair-and-why-should-i-care	4	3	0	0	0.431	0.077	8408046	EnergyEast	NatObserver	achievem… worth, benevolenc… care,
Oct-05-2017	Mike's excellent reporting on #EnergyEast provided citizens with vital, accurate, fair, tough information. Thanks, @mikedesouza [illegible]		0	1	0	0	0.018	0	54232308	EnergyEast	mikedesouza	universalis…
Oct-05-2017	US environmentalists and big oil have united to kill #EnergyEast . [illegible]		0	1	0	0	4.545	0	4554	EnergyEast		achie…

Page 14 of 98 10 Rows 131 - 140 of 977

Figure 6. Tweets and Users Table

To examine the influence of the users' tweets, we performed a statistical analysis of the engagement rate that includes likes, retweets, quotes, and replies from each individual tweet per user. As a result, we obtained an average of 0.527, a median of 0.018 and a standard deviation of 14.637. As it can be noted, the standard deviation looks sparse, the reason for this could be because of irregularities between users, immensely high retweets, replies, quotes, likes, followers, and followings per individual tweet can significantly alter the cumulative engagement.

To conclude, based on the results that we obtained from the above visualizations, we were able to infer that this day was an event that was fairly equilibrated between supporters and opposers. People showed through their messages their discontent about such event. The conversation revolved around concepts related to employment, worthiness and what was the best thing to do. By looking up online through different news sources we were able to find out that on this day the Energy East pipeline was canceled by TransCanada.

The analysis of this dataset was done in a Linux PC with 16 cores - 56GB of RAM and took approximately 2 minutes to complete all the processing and make the data available online.

4.2 The Academic Institution Case Study

The second use case consisted of gathering the user's timeline of all the accounts that are officially associated with the University of Alberta (UofA), Canada. This was done through a client based on the Twint scraper, to scrape users' timeline. The objective was to help the communications department to identify accounts that should be terminated, and provide them with data on which to base their decisions.

We used our custom table for exploratory analysis to cover the following sections:

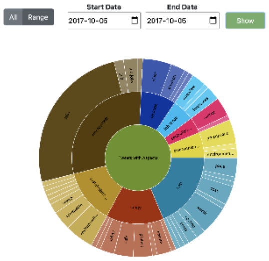


Figure 7. Sunburst Chart for Frequency of Aspect Terms

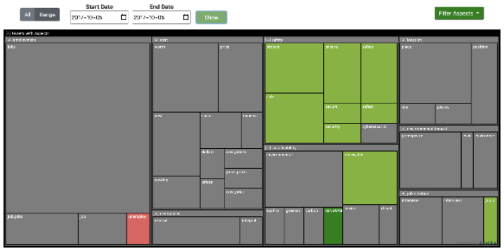


Figure 8. Treemap for Frequency of Aspect Sentiments and Terms

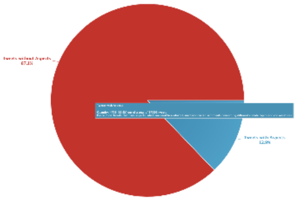


Figure 9. Pie Chart of Tweets without Aspects vs Tweets with Aspects

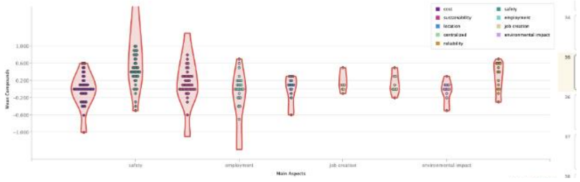


Figure 10. Violin Chart for Distribution and Density of Sentiment Compounds per Aspect

- Tweets: Accumulation of all the tweets from the official university's accounts.
- UofA accounts: Individual accounts from the university.
- Other accounts tweets: Accumulation of all the tweets from the interested external accounts.
- Other accounts: Individual accounts external to the university's accounts.

The engagement rate indicates how much users are interacting with the content; the extended reach metric captures the rate of tweets distribution to reach its audience; and the potential impressions metric captures how many users the content could impact. For each tweets table a statistical analysis was performed based on the engagement of all the tweets. And for each accounts table the same metrics were calculated but in an individual way to see their distribution.

This case study differs from the previous one, where a hashtag was the main study to calculate engagement. When it is about a brand, all the tweets that are fetched are about the accounts of interest. In this case we obtained an average of 0.124, a median of 0.007 and a standard deviation of 0.669. The standard deviation is fairly sparse, and we notice that there is a big variability since the median is small. The next step would be to filter by individual account and see their individual engagement rate, compare it against the cumulative, and see if it is just an outlier that needs to be terminated or requires more content to make it more popular. As an example, Table 1 shows the top 10 engaging tweets' ids from the main UofA account with their respective engagement rate, further analysis needs to be done to propose a strategy to increase new posts' engagement.

Table 1. Top 10 Engaging Tweets' Ids from the UofA

1216404327083823104 3.634		
1329884789646749699 2.541	1295734003404701696 0.804	1214973079924789248 0.667
1313111958078410753 2.484	1238353581419270144 0.69	1215392396713709568 0.611
826882183365328896 1.215	1214950781742272512 0.678	861959050757591041 0.568

5 Discussion and Conclusions

Data visualizations are especially useful to illustrate complex conversations, across social media platforms, to multiple user groups. To meet this challenge, we have developed a robust and scalable platform for the analysis of three semantic aspects of tweets - personal values, sentiment, and humor expressed in them – and the public's engagement to measure influence. Lastly, we develop complementary visualizations to enable users to gain insights on these data. We demonstrate the usefulness of our platform with two case studies. Our platform is scalable and allows handling higher workloads via scale out. Moreover, it is also extensible: it is implemented with clear APIs allowing the addition of new and more sophisticated analysis.

There is still potential for providing more details in the aforementioned charts, as well as adding more visualizations. This includes adding more aspects - as currently there are only nine -, more aspect terms and term combinations, and possibly using other visualizations like stream graphs that may be suitable for relationships like that of personal values, sentiment, and humor.

ACKNOWLEDGMENTS

We sincerely acknowledge funding from David Milia and the Canadian Energy and Climate Nexus, Mitacs, and AI4Society at the University of Alberta, Canada.

REFERENCES

[1] Alan S. Abrahams, Weiguo Fan, G. Alan Wang, Zhongju John Zhang, and Jian Jiao. 2015. An integrated text analytic framework for product defect discovery. Prod. Oper. Manag. 24, 6 (2015), 975–990.

[2] Bouktaib Adil, Fennan Abdelhadi, Bahra Mohamed, and Hmami Haytam. 2019. A spark based big data analytics framework for competitive intelligence. In 2019 1st International Conference on Smart Systems and Data Science (ICSSD), IEEE, 1–6.

[3] Santa Agreste, Pasquale De Meo, Emilio Ferrara, Sebastiano Piccolo, and Alessandro Provetti. 2015. Trust networks: Topology, dynamics, and measurements. IEEE Internet Comput. 19, 6 (2015), 26–35.

[4] Hager Ahmed, Eman M. G. Younis, Abdeltawab Hendawi, and Abdelmgeid A. Ali. 2020. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Gener. Comput. Syst. 111, (2020), 714–722.

[5] Sultan M. Al-Daihani and Alan Abrahams. 2018. Analysis of academic libraries' Facebook posts: Text and data analyt-ics. J. Acad. Libr 44, 2, (2018), 216–225.

[6] Ebtesam Alomari and Rashid Mehmood. 2018. Analysis of tweets in Arabic language for detection of road traffic conditions. In Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer International Publishing, Cham, 98–110.

[7] Ebtesam Alomari, Rashid Mehmood, and Iyad Katib. 2019. Road traffic event detection using twitter data, machine learning, and Apache spark. In 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), IEEE.

[8] Chester I. Barnard. 1938. Functions of the executive (30th ed.). Harvard University Press, London, England.

[9] Jurafsky Daniel and James H Martin. 2020. Dependency Parsing. InSpeech and Language Processing(3rd draft ed.). Chapter 14.

[10] Mary Douglas. 1966. Purity and danger. Routledge, New York.

[11] Joel Gehman, Linda K. Treviño, and Raghu Garud. 2013. Values work: A process study of the emergence and performance of organizational values practices. Acad. Manage. J 56, 1, (2013), 84–112.

[12] Wu He, Shenghua Zha, and Ling Li. 2013. Social media competitive analysis and text mining: A case study in the pizza industry. Int. J. Inf. Manage. 33, 3 (2013), 464–472.

[13] Steven Hitlin. 2003. Values as the core of personal identity: Drawing links between two theories of self. Social Psychology. Q. 66, 2 (2003), 118.

[14] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy:Industrial-strengthNaturalLanguageProcessinginPython. https://doi.org/ 10.5281/zenodo.1212303

[15] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, Seattle, Washington, USA.

[16] C. Hutto and E. Gilbert. 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.

[17] M. E. J. 2003. Newman The structure and function of complex networks. SIAM Rev 45, 2, (2003), 167–256.

[18] Richard W. Kilby. 1993. The study of human values. University Press of America, Lanham, MD.

[19] C. Kluckhohn. 1951. Values and value-orientations in the theory of action: an exploration in definition and classification. In T. Parsons and E. Shils (eds.). Harvard University Press, Cambridge, MA, 388–433.

[20] Peter M. Landwehr, Wei Wei, Michael Kowalchuck, and Kathleen M. Carley. 2016. Using tweets to support disaster planning, warning and response. Saf. Sci. 90, (2016), 33–47.

[21] M. D. P. Lee and M. Lounsbury. 2015. Filtering institutional logics: Community logic variation and differential responses to the institutional complexity of toxic waste. Organization Science 26, (2015), 847–866.

[22] L. M. Lefsrud, C. Westbury, J. Keith, and G. Hollis. 2015. A Basis for Genuine Dialogue: Developing a Science-Based Understanding of Public/Industry Communication.

[23] Lianne M. Lefsrud and Renate E. Meyer. 2012. Science or science fiction? Professionals' discursive construction of climate change. Organ. stud. 33, 11 (2012), 1477–1506.

[24] Lianne Lefsrud and Achim Oberg. 2019. Heated atmosphere: Organizational emotions and field structuring in online climate change debates. Acad. Manag. Proc. 2019, 1 (2019), 12132.

[25] Vittorio Lingiardi, Nicola Carone, Giovanni Semeraro, Cataldo Musto, Marilisa D'Amico, and Silvia Brena. 2020. Mapping Twitter hate speech towards social and sexual minorities: a lexicon-based approach to semantic content analysis. Behav. Inf. Technol. 39, 7 (2020), 711–721.

[26] Duncan McCaig, Sudeep Bhatia, Mark T. Elliott, Lukasz Walasek, and Caroline Meyer. 2019. Analysis of Twitter messages using big data tools to evaluate and locate the activity in the city of Valencia (Spain) Cities 86, (2019), 37–50.

[27] Duncan McCaig, Sudeep Bhatia, Mark T. Elliott, Lukasz Walasek, and Caroline Meyer. 2018. Text-mining as a methodology to assess eating disorder-relevant factors: Comparing mentions of fitness tracking technology across

online communities. *Int. J. Eat. Disord.* 51, 7 (2018), 647–655.

[28] R. Mechler. 2016. Reviewing estimates of the economic ef-ficiency of disaster risk management: opportunities and limitations of using risk-based cost–benefit analysis. Nat. Hazards (Dordr 81, 3, (2016), 2121–2147.

[29] W. Medhat, A. Hassan, and H. Korashy. 2014. Sentiment analysis algorithms and applications: A survey, Ain Shams Eng. J 5, 4, (2014), 1093–1113.

[30] Cataldo Musto, Giovanni Semeraro, Pasquale Lops, and Marco de Gemmis. 2015. CrowdPulse: A framework for real-time semantic analysis of social streams. Inf. Syst. 54, (2015), 127–146.

[31] Evgenii Nikitin. RedditScore. https://github.com/crazyfrogspb/RedditScore

[32] A. Oberg, L. M. Lefsrud, and R. A. Meyer. 2021. Organizational (Issue) Field Perspective on Climate Change. Economic Sociology: The European Electronic Newsletter (2021).

[33] N. Panagiotou, I. Katakis, and D. 2016. Gunopulos Detecting events in online social networks: definitions, Trends Chall.

[34] Viktor Pekar, Naveed Afzal, and Bernd Bohnet. 2014. UBham: Lexical resources and dependency parsing for aspect-based sentiment analysis. Strouds-burg, PA, USA.

[35] Thomas J. Peters and Robert H. Waterman. 1982. In search of excellence. Harper & Row, New York.

[36] Michal Podhoranyi and Lukas Vojacek. 2019. Social media data processing infrastructure by using Apache spark big data platform: Twitter data analysis. In Proceedings of the 2019 4th International Conference on Cloud Computing and Internet of Things - CCIOT 2019, ACM Press, New York, New York, USA.

[37] Vladimir Ponizovskiy, Murat Ardag, Lusine Grigoryan, Ryan Boyd, Henrik Dobewall, and Peter Holtz. 2020. Development and validation of the Personal Values Dictionary: A theory–driven tool for investigating references to basic human values in text. Eur. J. Pers 34, 5, (2020), 885–902.

[38] Anuja Prakash Jain and Padma Dandannavar. 2016. Text analytics framework using Apache spark and combination of lexical and machine learning techniques. J. Appl. Inf. Sci. 4, 1 (2016), 31–36.

[39] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Stroudsburg, PA, USA.

[40] Milton Rokeach. 1973. The nature of human values. Free Press, New York, NY.

[41] Androniki Sapountzi and Kostas E. Psannis. 2018. Social networking data analysis tools & challenges. Future Gener. Comput. Syst. 86, (2018), 893–913.

[42] S. H. Schwartz. 1992. Universals in the content and structure of values. In M. P. Zanna (ed.). Academic, San Diego, 1–65.

[43] S. H. Schwartz. 1994. Are There Universal Aspects in the Structure and Contents of Human.

[44] S. H. Schwartz. 1999. A theory of cultural values and some implications for work. Applied Psychology 48, (1999), 23–47.

[45] P. Selznick. 1957. Leadership in administration: A sociological interpretation. Harper & Row, New York.

[46] Robyn Speer, Joshua Chin, and Catherine Havasi. 2016. ConceptNet 5.5: An open multilingual graph of general knowledge.

[47] Guilherme M. Thomaz, Alexandre A. Biz, Eduardo M. Bettoni, Luiz Mendes-Filho, and Dimitrios Buhalis. 2017. Content mining framework in social media: A FIFA world cup 2014 case analysis. Inf. manag. 54, 6 (2017), 786–801.

[48] Andrew H. Van Ven, Thomas J. Peters, and Robert H. Waterman. 1983. In search of excellence: Lessons from America's best-run companies. Adm. Sci. Q 28, 4, (1983), 621.

[49] Nathan Walter, Michael J. Cody, Larry Zhiming Xu, and Sheila T. Murphy. 2018. A priest, a rabbi, and a minister walk into a bar: A meta-analysis of humor effects on persuasion. Hum. Commun. Res. 44, 4 (2018), 343–373.

[50] Max Weber. 1905/2002. The protestant ethic and the spirit of capitalism with other writings on the rise of the west (4th ed.). Penguin, New York.

[51] Chris Westbury and Geoff Hollis. 2019. Wriggly, squiffy, lummox, and boobs: What makes some words funny? J. Exp. Psychol. Gen. 148, 1 (2019), 97–123.

[52] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 - EMNLP '09, Association for Computational Linguistics, Morristown, NJ, USA.