University of Arkansas – CSCE Department
Capstone I – Final Proposal – Fall 2021

Walmart: Social Media Analysis

Daniel Salazar, Jonathan Montoya, Kayla Hernandez, Joshua Shackleton, Tanner Mecham, Caleb Joiner

Abstract

As the number one Fortune 100 company, Walmart has 5,000+ stores in the United States and nearly 2.4 million associates throughout the world. In order to meet their wants and needs, Walmart continuously attempts to implement methods to gauge associate satisfaction and intercompany sentiment. It is not uncommon for employees to be hesitant to share their dissatisfactions with their superiors, and to address this issue, Walmart aims to improve the understanding of associate sentiment by analyzing social media trends to assist in better understanding their associates thoughts & feelings on the company and its decisions. By accomplishing this, Walmart can be better suited to make more associate-tailored decisions & more accurately understand the thoughts, feelings, and wishes of all Walmart associates.

1.0 Problem

Consumers across the globe depend on stores to have the products that they want. Without stores that supply the needs of the everyday consumer life would be a bit more barbaric. Walmart is in the industry of meeting the needs of individuals, in the sense that they provide a product for money. In other words, Walmart is a retail store. The problem is that being a retail industry, and one of the largest ones in the world, they have to keep their employees happy and productive.

Deciding which stores need certain things for their employees may be difficult, especially from a corporate-level. These needs of the employees could depend on the demographic area to simply how their managers manage. The problem that Walmart faces is that there are many, many different individuals at Walmart that all have different needs and see different things that can be done to make Walmart an overall better place, understanding each and every individual's concerns is difficult.

2.0 Objective

Large corporations such as Walmart are notorious for collecting data to understand user / consumer sentiment and use this data to quantify their efforts, profits, potential, sales, and other metrics. These are all relatively easy to capture, and are all geared towards providing a service that is of utmost benefit to the consumer. The objective of this project, however, is not geared

towards better understanding the market or consumer, rather to aid Walmart management to better understand the internal company sentiment and how associates address and discuss issues and successes in the workplace with those outside of their internal peer or management group. The end goal of this project is to provide web-scraping software & analysis tools that will allow Walmart to collect an anonymized dataset on publicly-posted associate speech through web data mining. This in turn will help the company better understand how their employees feel about intercompany changes, procedures, policies, etc., through gauging user sentiment and thorough analysis on use of keywords, phrases, and other indicators. This information can then be used by company leadership to make changes in intercompany policy or make decisions that will have a higher net positive effect and impact on the employees of Walmart.

3.0 Background

3.1 Key Concepts

Each day, nearly 2.5 quintillion bytes of data are generated due to web usage. Much of this data is user-uploaded to social media platforms like Twitter, Instagram, Facebook and Reddit, and often includes keywords that can be traced back to a certain issue, product, service, or other area of discussion. The goal of this project is to develop verbiage analysis and information scraping tools to analyze the frequency and context of these keywords through mass-exporting text and other metadata (posting time, location, post interactions, etc.) from these sources. Using this data, the intention is to implement a method to analyze and search for particular keywords to understand employee sentiment (good/bad usage) on issues pertaining to the employees of Walmart to help company management better understand the unspoken or unreported impact company changes, policies, or decisions may have on its employees.

3.2 Related Work

There have been many similar implimentations of data analysis from web scraped items on social media. Social media is a great way to gather data given it users use it solely for their own enjoyment and purposes rather than work related or for some other reason. Therefore the data that is scraped off of these medias are a good indicator of what a natural response to different products and trends in general are.

One example of this would be pizza businesses promoting their products on social media and pulling data from the posts to learn more about what products are best. This has become such a large practice for pizza joints that “85% of pizza-chain sales are now tied to promotions and discounts mostly acquired through social media accounts” [1]. While this works well for pizza chains, Walmart’s implementation of this will have to go a step further. They will need to not only provide advertisements and look at the sentiment to find which products are more liked, but it will also require the ability to compare multiple different types of products as well as the number of those products that need to be stocked for a given location. There are not necessarily problems with the pizza joints method to use social media, but we are improving on the overall method by looking at what walmarts customers want from the company and how we can find a solution to it with the given data.

A report on data mining techniques [2] talks about how they would sort data based on clustering. One specific method of clustering is partitioning. This is a great way to contain data in specific

manners that would allow ease of access. For this specific project we would be able to implement clusters and have the clusters interact with each other in different ways. Each cluster would be a group of walmarts that all have similar geographic locations, or economic areas. That way inside the clusters we have specific needs, and we can use clusters that are next to each other to have similar but not absolute data. One tool that will be necessary in order to complete this task is to realize that we will be taking in unstructured data. So we can't easily store a string or an integer of X length. We have to handle unknown data which "allows for the capture and integration of customer dialog, sentiment, and agent performance" [3].

4.0 Design

4.1 Requirements and Design Goals

Design Goals: (what do we need to do in order to make our system work)

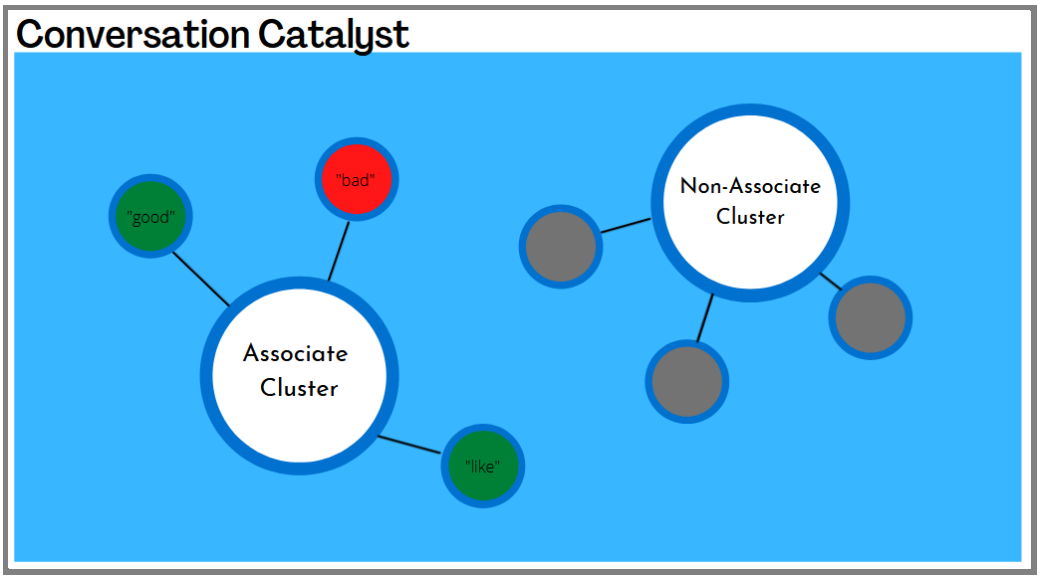
- Find a method to identify conversation catalysts, which can be personalities or events that induce discussion about certain topics
    - Subset catalyst: Walmart announcements, seeing how associates react given internal changes
- Create "associate clusters", defined groups of associates that we can further analyze into "pro/against" sentimental arguments
    - Can use keyword analysis to define group, or simply assume such a group exists
- Use conversation catalysts and associate clusters to identify associate sentiment about a given topic

Secondary Goal:

- Given a stream of data, identify the conversation catalysts / cluster-like traits that define topics and lead to sentiment

Requirements:

- Formulate a methodology of classifying associates (associate clusters) and utilize it on data sets to only gain insight from associates
- Create an appropriate modelling tool that satisfies design goals
- Integrate tool into native Walmart applications, for future use
- Form a comprehensive, automated process around the tool so that no associate work is necessary for automatically analyzed sentiments.



4.2 Design - System Architecture

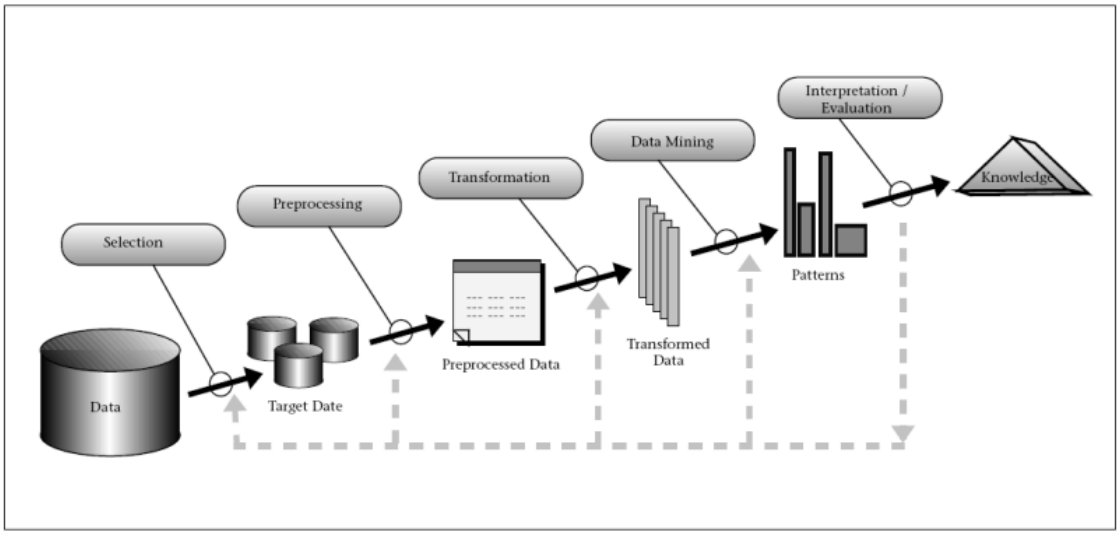
While social media sentiment analysis is not a new idea, in designing this system we have to answer a few questions; firstly, how do we determine if the user is a Walmart associate? Then, what information are we trying to predict using our machine learning model that we can weigh against associate sentiment information to determine accuracy? Lastly, how accurate do we need the model to be? We will use some ideas from previous implementations of social media sentiment analyses that some team members have worked on, and then build upon that model to work for our purposes (Walmart -> Twitter attitudes toward Walmart operations. etc.). Once this is done, we will refine the prediction model to meet or exceed some arbitrary threshold that will be defined in the project requirements (we can say 85% accuracy for now). The idea goes as follows:

Design Goals:

- Form research on which keywords / catalysts we should be looking for when deciding if we see social media sentiments pointing to Walmart activities
- Create a method of identifying or classifying a subset of individuals as associates
- Based on these keywords and classifications, scrape data off social media sites such as Twitter, Facebook, and Reddit in order to gain insight into associate sentiment about events / products / etc.
- Feed this data through a machine learning model and train the model to create groupings: good feedback, negative feedback.
- Test this machine learning model against a known data set pertaining to previous sentiments from associates, by creating a test set of data.
- Refine the prediction model in order to get it to an acceptable accuracy, and utilize the model to predict and analyze associate sentiment in the future.

Model Goals:

- Use an unsupervised learning model to discover patterns in text entries containing posts/comments from social media platforms
- Model should be able to get insights from large volumes of new data
- Format should contain a small text entry (character limited) and some information to collect insights such as whether or not the person is a Walmart associate, whether or not the information is related to a catalyst, and what region/location the post is from
- Use the K-means algorithm to cluster the data
- Training on data sets will include keywords, Natural Language Processing based learning, and information retrieval techniques
    - Character-limited entries helps us limit refinement needed to training sets
- Manual insight validation to determine acceptable results



4.3 Risks

Risk	Risk Reduction
Exposing sensitive user data	Ensure everything is captured with anonymity (name fields fully purged etc)
Biased scraping; not portraying accurate trends = skewed data	Pull from multiple sources of different demographics to get a full unbiased picture of general trends
Incorrect associate clustering	Training methodologies such as manual review to guarantee model learns and avoids issue

4.4 Tasks

1. Clearly define requirements of the project as set by Walmart sponsor and brainstorm ideas on how we will fulfill those requirements.
2. Review requirements with sponsor and team to ensure they are fully understood by both parties.
3. Create low-level design for the entire project and review design with the team. including all drawbacks and tradeoffs. Evaluate risks, runtime, memory, security, and other metrics.
4. Implement the designed system in an agile environment to maintain and extend code from iteration to iteration to clarify and simplify design without changing its behavior.
5. Write unit tests and review code regularly to maintain quality to ensure it meets industry standards. perform sanity checks to minimize regression. Perform sanity checks regularly to minimize regression.
6. Write a document describing code for users and developers that describes problems solved, including small code samples with a link to the code and issue tracker.

4.5 Schedule

Tasks	Dates
1. Clearly define requirements of the project set by sponsor and brainstorm ideas on how to fulfill requirements.	11/1-11/8
2. Review requirements with the entire team to ensure they are fully understood by both parties.	1/18-1/25
3. Create low-level design for our primary and secondary goals and review design with the team.	1/26-2/7
4. Implement the designed system in an agile environment to maintain and extend code.	2/8-3/22
5. Write unit tests and review code regularly to maintain quality to ensure it meets industry standards.	3/23-4/6
6. Write documentation describing code for users and developers that describes problems solved.	4/7-4/14

4.6 Deliverables –

- Project Document: Contains a thorough description of requirements set by the team. This includes system code and timeline of meetings, notes, and sing-offs.
- Database schema and data: This includes data used in our project that we have collected from third parties or in-house.
- Code: This includes the code for the system along with comments, naming conventions, and programming practices.

- Final Report: A comprehensive report outlining a project in detail, specifically pertaining to planning phases, development phase, iterations of codes, trade-offs, results, actual sprint time-line in comparison to planned time-line.
- Final Deliverable - all previous reports and created code for the project

5.0 Key Personnel

Caleb Joiner - is a senior Computer Engineering major in the Computer Science and Computer Engineering department at the University of Arkansas. He has industry experience at J.B. Hunt as an applications development intern working on full stack angular and spring boot applications as well as web automation applications in javascript which included the concept of scraping data off of a website. In school the most applicable class he has taken was Information retrieval which consists of weighting terms and manipulating data to help come up with the best results of a search, similarly to how google works. He will work with the team to apply his web scraping skills as well as Information Retrieval skills to help analyze social media data for the application.

Daniel Salazar - A senior Computer Science major in the Computer Science and Computer Engineering department at the University of Arkansas. He has industry experience working as a Software Development Engineer at Amazon and has also contributed to many open-source projects. He has completed sentiment analyses in Big Data and also has relevant experience in Information Retrieval. He will be responsible for validating the design of the system in terms of efficiency, scalability, and feasibility. He will also be responsible for writing Python code for the system and reviewing code commits from others to ensure all of the code is high quality, extensible, and up to industry standards.

Jonathan Montoya - A senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has taken a Big Data and Analytics course which dealt with collecting large amounts of data and finding trends within them. He created a project that collected Tweets from Twitter and stored them into a database and ran queries to find certain patterns. Tools used in this project included twitter API key, MongoDB, and python sentiment analysis libraries. He will work on the design team of the system.

Joshua Shackleton - is a senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has completed Big Data Management and Artificial Intelligence, both classes which focus on creating algorithms to analyze large amounts of data and utilizing machine learning to predict trends in data. He has also been a part of a team of students at the McMillon Innovation Studio working with Walmart to improve the returns process within Walmart stores worldwide. He will work to assist his fellow design team members in creating and training the algorithms that will be used in the project.

Kayla Hernandez - A senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. She has gained industry experience while interning with J.B. Hunt Transport Services, Inc., where she has been interning for nearly 3 years. She has primarily worked with the Angular framework on various applications prepared for use internally within the company. She has not had much experience in any machine learning

related work, so she will be responsible for assisting in organizing the data pulled, but mostly learning about NLP and sentiment analysis.

Tanner Mecham - A senior Computer Science major in the Computer Science and Computer Engineering Department at the University of Arkansas. He has worked to develop and maintain professional relationships with Walmart through creation and design of network cabling pre-build plans during his last 3 years at Wachter, Inc. He has undergone studies in the fields of AI, cloud data processing, and network architecture, and is proficient in Python & Java. He will be responsible for developing visualizations of the methods, tools, and data created/collected.

Anton Groom - is a Senior Director of Digital Transformation, Enterprise Business Services at Walmart with a focus on delivering digital advantage with AI products and Automation services that optimize the Human-Machine team to drive efficiencies and empower people to achieve their aspirational goals

David Daniel - is a Walmart employee working under Anton with data analytics and machine learning in the RADIAL-7 team.

Kevin Horecka - is the principal developer of the RADIAL-7 team at Walmart. He has a background in computer science, neuroscience, and natural language processing, and has been at Walmart for three years.

6.0 Facilities and Equipment

We will be performing sentiment analysis on Reddit as part of our final project. Reddit's API is somewhat flexible; one can easily find data that interests them by focusing on a specific subreddit. The Walmart sponsors for this project have expressed why Reddit is where they have chosen to grab data to be able to best understand their associates: it contains subreddits of users that self-identify as Walmart associates, so, it would almost guarantee that the sentiment analysis being performed is accurate and targeting the correct group of users. Although Reddit contains millions of different subreddits, our team could find some of the most active ones of users self identifying as Walmart associates and perform the sentiment analysis on those subreddits.

To be able to perform sentiment analysis on Reddit, we will need to register an app through Reddit. This would give us access to the client ID, client secret and user agent – all necessary to give us access to Reddit's data and fetch posts. Python has several easy-to-use libraries to be able to perform sentiment analysis, so it is our language of choice for this project. A few examples of these libraries our team would leverage are praw - a Python library used for accessing the official Reddit API - emoji - a Python library used to remove emojis from a post - and re - a Python library used to remove links from a post. There are other Python libraries that we may choose to use that could help us with our project, such as pandas, which would aid in organizing the data pulled to better observe trends.

7.0 References

[1] Social media competitive analysis and text mining: A case study in the pizza industry, https://www.sciencedirect.com/science/article/pii/S0268401213000030?casa_token=4hXIeRi8_KoAAAAA:Y4-m5s1R7Reh47omInhk_dMIVWCNA6RIZ8lad6AxckgUS8IlWat13Wibn2_T5zF-5DdOq6ek8w#bib0135

[2] DATA MINING TECHNIQUES AND APPLICATIONS, https://www.researchgate.net/publication/49616224_Data_mining_techniques_and_applications

[3] Data Mining: Definition, Techniques, Tools & Tips, https://callminer.com/blog/what-is-data-mining-definition-techniques-tools-and-tips-from-experts