**Tweet Sentiment Extraction in Business Applications**

Filipp Krasovsky

University of San Diego

Master of Science, Applied Data Science

ADS 501

Section 2

1-18-2021

# Background

Customer-facing brands in a variety of industries (retail, customer service, banking, etc.) face a large amount of uncertainty on a day-to-day basis in determining how their customers feel about them, holistically, as an entity. Public perception can be incredibly volatile and subject to change based on the news cycle, actions taken by the PR team, and more. A recent example includes the spotlighting of Twisted Tea, a beverage produced by the Boston Beer Company, after a viral YouTube video displayed an altercation between two people in which one was struck over the head with a bottle of the beverage. Calls for the company to sponsor the individual who used Boston Beer’s product as a weapon took center stage in the media cycle, creating space for the company to leverage media attention to drive sales. (Newsweek, 2020)

Consequently, an understanding of public perception is crucial to a corporation in many dimensions; a lack of identifiable public perception creates a call to action for a company to expand its marketing efforts, while an overwhelmingly positive public perception serves as either a proxy for the success of a campaign or an indicator of diminishing marginal returns on marketing – if public perception is already great, less value is generated per dollar spent on every ad campaign. Similarly, negative overall perception can create a call to action to create sales promotions to salvage a customer base. For instance, being able to track an individual customer’s sentiment towards the company informs future customer service calls and can provide the firm with an understanding of what kind of treatment they need to give in order to maximize their chances of retention (Forbes, 2020).

The problem companies face is being able to engage both potential and current customers to understand how they feel about the brand during a given moment in time. Polling is both costly and ineffective, often creating biased samples while failing to reward customers for filling out surveys (Forbes, 2019). More importantly, polling is often product-oriented and can fail to consider outside events that may influence the customer experience, such as the news scandal that placed Target at the center of a series of accusations of spying on their customer base after it was alleged that their predictive analytics program “uncovered” a teenage pregnancy before it happened (Lubin, 2012).

In this proposal, we set the context of our organization as being a well-known, nationwide retailer – analogous entities could include Ford, Urban Outfitters, The Boston Beer Company, etc. Breaking the fourth wall momentarily, we observe the key business problem described in the Kaggle dataset chosen for this project:

*“With all of the tweets circulating every second it is hard to tell whether the sentiment behind a specific tweet will impact a company, or a person's, brand for being viral (positive), or devastate profit because it strikes a negative tone.”*

*(Kaggle, Tweet Sentiment Extraction, 2020)*

In essence, our organization has a substantial media presence and is a customer-facing business, most likely rooted in the retail space. The key challenge we aim to address is, provided a considerable amount of “noise” about the company circulates throughout different communication channels, whether it is possible to tell how much of the publicity the brand receives is positive or negative. We assume this corporation as a key stakeholder for this project; the results of this project are actionable for both its leadership as well as its customer relations personnel. More specifically, it is actionable for the following entities in both cases where sentiment is found to be increasingly positive or increasingly negative:

|  |  |
| --- | --- |
| Department | Application |
| Marketing/Advertising | If sentiment deviates to be more negative, the analytics product encourages advertising to tie brand to socially beneficial causes and charities to increase positive perception and boost sales in aggregate. If sentiment deviates to be more positive, product encourages allocating less money to marketing campaigns. If sentiment is largely neutral or lacks critical mass, provides call to action to increase brand recognition campaigns. |
| Customer Service | Individual customer sentiment determines the customer service department’s retention strategy – specifically, how many discounts, promotions, gift cards, etc. need to be given to a customer to retain their business. |
| Public Relations | Aggregate increases in negative and positive sentiment can be used to determine the effect of the news cycle on company profits – negative sentiment is a call to action to use media as a platform to undo negative perceptions about company image, such as going on a news network to discuss a controversy related to the business. |

Table of Contents

[Background 2](#_Toc61898487)

[Inventory of Resources 6](#_Toc61898488)

[Terminology 7](#_Toc61898489)

[Initial Data Collection Report 8](#_Toc61898490)

[References 9](#_Toc61898491)

# Inventory of Resources

**Data Source:** A series of Twitter Posts (Tweets) aggregated under the creative commons license.

**Legal Personnel:** The company retains at least one attorney knowledgeable about intellectual property – this individual will help us determine the usage of data compiled through a creative commons license in a profit-seeking environment. Furthermore, the company has a legal subject matter expert who can determine if the usage of Tweets violates any regional or federal privacy laws.

**Business Development Personnel:** These individuals will provide the main feedback on the UX for the analytics deliverable to make sure that there are optimized variants for the customer service and marketing teams, as well as executive leadership.

**Software/Data Access:** Data can be retrieved remotely from the Kaggle API in the form of three CSV files – alternatively, given that the sample for this project is static, the company can store this data internally on a NoSQL database.

**IT/Software Development Personnel:** This group is responsible for integration of the analytics product into existing company software, maintenance of data integrity and storage, and troubleshooting/refactoring during the evaluation and modelling process.

# Terminology

1. Tweet – refers to a single post made by an individual on Twitter.
2. NoSQL Database – a non-relational database that doesn’t require an explicit schema upfront in order to operate, allowing for horizontal scalability.
3. Customer Retention – refers to the general practice of optimizing a strategy to ensure that existing customers continue engaging with the company.
4. Sentiment analysis – the technical process of parsing through text to determine whether the writer or author is displaying an overall positive or negative mood. This is often joined with identifying a subject matter to determine if the post expresses sentiment about an interested party, such as a corporation.
5. Polarity – refers to whether a body of natural language is positive, negative, or neutral. This can be expanded to more complex emotions such as happy, angry, and sad.

# Initial Data Collection Report

Our dataset consists of three different CSV files – a training dataset (n~27400), a testing dataset (n~3500), and a sample submission dataset derived from the testing dataset (n~3500). For the purposes of this project, the third dataset is tentatively irrelevant in the end product, but serve as a useful domain for evaluating performance. The dimensions for the testing and training datasets are as follows:

|  |  |  |
| --- | --- | --- |
| Field | Description | Used In |
| textID | Unique Identifier for a tweet. This is our **primary key.** | Training set, Testing Set, Sample Submission |
| text | The literal text of the tweet | Training set, Testing set |
| Selected\_text | The portion of the text which is used to identify the polarity/sentiment. | Training set, Testing set, Sample Submission |
| Sentiment | One of three possible overall moods characterizing the text – positive, negative, or neutral. | Training set, testing set |

**Additional Observations**

A face-value inspection of the raw data suggests an unusually high number of cases where the **text** and **selected\_text** were identical or approximately identical in cases where the sentiment was neutral. To confirm this, we placed the data in Excel and computed a derived value, **isEqual**, with the following logic:

*= (TRIM(B2)=TRIM(C2))*

Where B and C are columns housing the text and selected\_text dimensions, respectively.

Overall, cases where sentiment was neutral exhibit a ~90% likelihood that the selected text will be the same as the original text, while this occurs only 10% of the time for positive and negative sentiment identification.

Other possible challenges in identifying sentiment includes incoherent or non-natural word combinations, such as the presence of URLs in tweets, which cannot comprehensively be identified as good or bad, although technology exists to categorize the content of URLs themselves.

A final challenge is the observation of sarcastic data (Kaggle 2020) in the dataset, which might confound any attempt to understand the true sentiment of a tweet, such as “Everybody hates me, lol!” – a tweet that can be interpreted as negative, but also exhibits a propensity to be light-hearted and positive. Similarly, and more applicable to our organization, tweets such as “I just LOVE when Target jacks up prices!” run the risk of being interpreted as positive, even though they clearly exhibit hostility towards the firm.

Assuming this edge case manifests itself often enough, it may mislead our analytics product to calculate a higher volume of positive sentiment. Early stage solutions to this problem may include the application of a natural language API to detect sarcasm or to calculate the probability a tweet might be sarcastic and factor that into the value-added from our deliverable at the end.

# References

Forbes (Dec. 13, 2019). *Are Customer Surveys Effective?* Retrieved January 18, 2021 from https://www.forbes.com/sites/quora/2019/12/13/are-customer-service-surveys-effective/

Gus Lubin (Feb. 16, 2012) *The Incredible Story of how Target Exposed a Teen Girl’s*

*Pregnancy,* Business Insider, retrieved January 18, 2021 from

https://www.businessinsider.com/the-incredible-story-of-how-target-exposed-a-teen-

girls-pregnancy-2012-2

Kaggle (Mar. 23, 2020). *Tweet Sentiment Extraction – Extract Support Phrases for Sentiment*

*Labels*, retrieved January 18, 2021 from https://www.kaggle.com/c/tweet-sentiment-extraction/overview

Newsweek (Dec. 29, 2020). *Twisted Tea Video Viewed Over 1 Million Times, Inspires Wave of Memes,* Retrieved January 18, 2021 from https://www.newsweek.com/twisted-tea-video-1-million-times-memes-1557664