# 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

Inventory of Resources ..... 6

Terminology ..... 7

Business Objectives & Success Criteria ..... 8

Risks & Contingencies ..... 10

Data Mining Goals & Success Criteria ..... 12

Initial Data Collection Report ..... 13

References ..... 23

## Inventory of Resources

**Data Source:** A series of Twitter Posts (Tweets) aggregated under the creative commons license, as well as the Twitter API to fetch future posts to analyse for sentiment.

**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. We may also require a sarcasm detection API as outlined in the risks/dependencies section, and the ParallelDots Sarcasm API for initial EDA and overview.

**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.
6. NLP/Natural Language Processing – the practice of leveraging technology to sift through written and spoken language created by humans, often used for voice recognition and text-to-speech, as well as sentiment analysis.

## Business Objectives & Success Criteria

Our primary motivation for engaging in sentiment analysis is to leverage natural language processing capabilities to evaluate customer attitude towards the company in several layers:

1. Aggregate – how the public feels about the company both at a terminal evaluation (present-day) and over time. The latter data can also provide insights into the effects of an advertising campaign, media cycle, etc.
2. Demographic – provided we can gather data which allows us to creation a structural relationship between customer segments and sentiment expressed digitally, we can also develop the insights expressed in #1 on a more granular scale where input from a specific sub-group is more valuable (i.e. the effect of a diaper marketing campaign as evaluated by parent shoppers).
3. Individual – sifting through customer reviews, natural language processing from phone calls, emails, and other one-on-one communication with the company allows us to determine the current mood of a customer and what steps the firm must take to keep their business.

As a result, our business goals include utilize these insights in the following ways:

1. Ability to leverage performance metrics based on sentiment (ie, being able to gauge the success of a marketing campaign by the increase in positive sentiment over time). Being able to use the analytics product as a dependency in other reports is a success criterion as it provides the company with cost-saving insights on the marginal utility of a variety of marketing and PR deliverables. On a low level, this can look like actionable ways to reallocate the company budget or create a way to hold middle/upper management accountable for decreasing performance in a way that's more robust than engaging in

costly customer surveys.

2. Increase in customer retention, particularly through the customer service funnel, where NLP can be applied to correspondence with customers. If this analytics deliverable works as intended, the company should be able to leverage it to close churn rate/customer attrition. On a low-level, this means utilizing the deliverable to tell when a customer is dissatisfied enough to offer them adequate compensation to retain them in a sales funnel.

3. In aggregate, optimizing "damage control" resources (discounts to unsatisfied customers, PR campaigns, marketing/sales campaigns) based on a more accurate capture of customer sentiment should lead to a reduction in annual expenditures, and thus, net profit. Assuming the firm currently operates with some uncertainty about how its customers view it, it likely attempts to mitigate risk by insuring itself with additional public perception projects, sometimes beyond an optimal rate. If the analytics product can reduce uncertainty, we should observe a drop in spending.

## Risks & Contingencies

Several possibilities threaten the integrity, and thus, the success of the analytics product in being able to create the business success criteria mentioned above.

1. Sarcasm – most natural language processing engines have difficulty determining whether a statement is truly negative or positive, and seemingly positive statements can be sarcastic (and thus, negative). In our application, we anticipate a more frequent occurrence of sarcasm where the former half of a statement is positive while the latter half is negative (Mousa, 2016). If sarcasm is prevalent enough in content relating to our brand, we may crucially mis-allocate capital within the firm based on a misconception of how the public truly perceives the firm due to inability to detect sarcasm; this risk is compounded by questions about the representation of our customer base on social media.

   **Solution:** During the evaluation/modelling phase, leveraging additional resources such as a sarcasm detection API can show us a picture of how frequently we can expect statements about brands in general to be sarcastic. It's important that we narrow in on sarcastic remarks about customer experiences, as they may occur at a different rate than general displays of sarcasm. If sarcastic sentiment occurs significantly enough, a sarcasm API may become a permanent fixture of our analytics product and re-tag content appropriately.

2. Representation – this is inherent to the fact that (a) Twitter may not be the primary point of discourse about the firm and its services and that (b) social media may be misrepresentative of the demographic composition of our customer base. It is conceivable that Twitter may capture sentiment from Millenials and Gen-Z shoppers about the company but may not provide as much insight about Gen-X or Baby Boomers due to the

possibility that those groups may use the platform less. (Pew Research Center, 2019)

**Solution:** It is largely impossible to fix this issue as it is an intrinsic component of social media for the time being – we can, however, take action on the output of the analytics product with caution, ensuring that actionable insights are applied to relevant demographics.

3. Botnets – Experts claim that anywhere from 5 to 15 percent (CloudFlare, 2021) of Twitter accounts are run by bots, an automated program used to engage in social media. In the 2016 election, 20% of political discourse online was generated by bots alone. If our analytics product captures, internalizes, and forecasts sentiment based on communication put forth by bots, the company risks one of two scenarios:
    a. If it becomes public knowledge that the company uses sentiment analysis to make high-level business decisions, individuals in possession of botnets can deliberately spam social media with negative messages about the firm, making the deliverable largely useless.
    b. In the event the product remains hidden from public knowledge, bots must still be discounted to ensure accuracy.

    **Solution:** In the case of Twitter in particular, certain proxies such as number of posts, activity, and engagement with other accounts as well as creation date can be useful indicators that an account is a bot. We may have to implement the **Twitter API** as a mainstay in the deliverable to facilitate this safeguard.

## Data Mining Goals & Success Criteria

Our primary goal in producing an analytical model is to maximize accurate recognition of positive and negative sentiment. In practical terms, positive sentiment will be our positive target, and we are interested in maximizing the number of True Positives and True Negatives. As a result, our success criteria are:

1. Optimize the process of finding the key words in a tweet by focusing on adjectives and nouns while ignoring URLs and other illegible strings and characters.
2. Correctly matching key identifiers to their proper polarity (Positive/Negative)
3. If possible, mapping aggregate sentiment about the firm as a time series to provide insight for other aspects of firm performance as well as track individual customers in the sales funnel.
4. Flag sarcastic content on the individual level in contexts like the one in #3 to advise customer service staff on best practices for dealing with a given customer.

## 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 |

## Data Description/ 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. **Using an Excel search function, we can identify 1575 instances of the keyword "http",** suggesting a considerable prevalence of URLs that may throw off our analytics model.

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.

**Initial Breakdown**

The training set can be broken down by the three types of polarity available:

| Positive | Negative | Neutral |
|---|---|---|
| 32% (8582) | 28% (7781) | 40% (1118) |

The data does happen to contain some traces of sarcastic content, and was traceable to varying degrees by the **ParallelDots API**, which was used at the time of this report for demonstration purposes only.

| **Text** | **Likelihood Sarcastic** |
|---|---|
| **you guys didn`t say hi or answer my questions yesterday but nice songs.** | 50.6% |
| **it`s nice to leave the office when the sun is still up** | 33% |
| **And I just love every little thing about you...** | 15% |
| **I love the smell of procrastination in the morning... oder so.** | 70% |

*(Source: ParallelDots Sarcasm API + Kaggle Training Dataset)*

Although not statistically significant, we can infer from randomly perusing the dataset that sarcastic Tweets take up at least some portion of our training set, and some text is potentially ambiguous, such as in the case of the third Tweet, which could arguably be both sarcastic and non-sarcastic when analysed manually. This initial overview presents the biggest challenge and impediment to our dataset: identifying sarcasm and sifting past URL strings and other non-insightful phrases.

## Data Quality Report and Exploration

Given that we are not working with any ratio, interval, or numerical variables, we can only apply the type of summary figures that are valid for categorical data – in this instance, tweets, selected text from tweets, and their polarity.

For this study, we will focus entirely on analysing the English language in social media. As a result, our first order of business is to assess how much of our dataset meets the criteria of being in English. Grammar also presents itself as a problem in that even leading software has trouble identifying sentiment for strings that are grammatically incorrect, use slang, or have irregular spellings (ex. **"soooooooooo tired!"**). Nevertheless, a high-level overview shows us that most of our ABT is aggregated from English social media content.

By utilizing the **CLD3** package in R, as well as a sequence of functions that removes unnecessary strings, urls, and grammatically confusing character sequences, we can get a rough approximation of how much of our ABT can be interpreted by an NLP model. We begin with a small cross-section of our data by getting the average frequency of English tweets in 10 samples of n=1000:

**Average English Classification of Tweets (text field, n=1000, sampled 10 times)**

| Language | Average Frequency |
|---|---|
| **English** | 0.7878 |
| **Non-English (mismatched)** | 0.2122 |

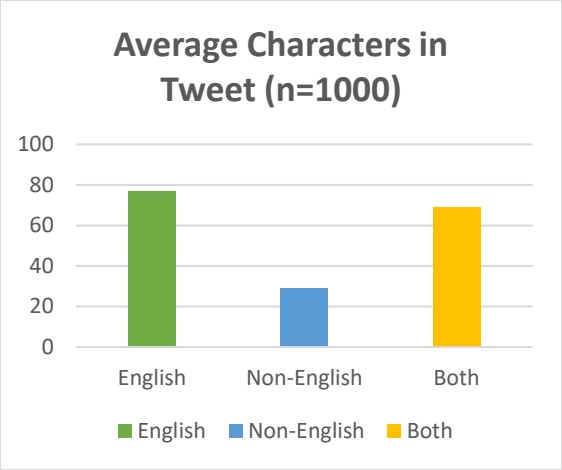
*Generated using the cld3 package in R 4.03*

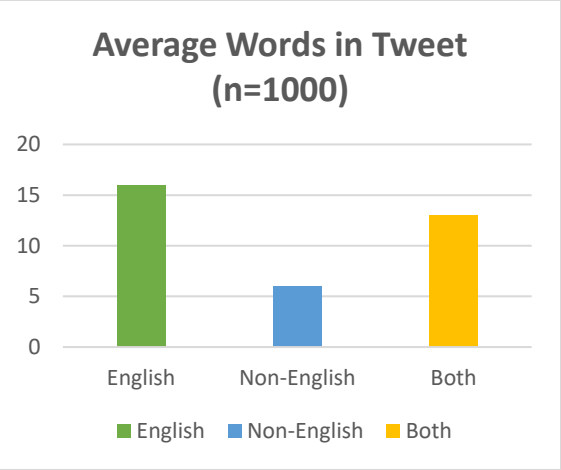
On face value, about 79% of our ABT can be readily moved into further modelling and exploration. Given that the dataset extracted for this project is very likely written entirely in

English, some investigation is required to determine the possible causes of poor data. We can validate the frequency of tweets classified as English ought to be much higher as we have domain expertise that the entire dataset is extracted from English-speaking Twitter pages.

We can pursue this train of thought by taking one of our samples (n=1000) and examining several factors that might contribute to the noise that prevents classification – in particular, the length of the tweet and whether it contains grammatical errors. In observing these factors, we are exploring a relationship between our ABT's misclassification rate and the features of its text as well as conducting a quality report at the same time.

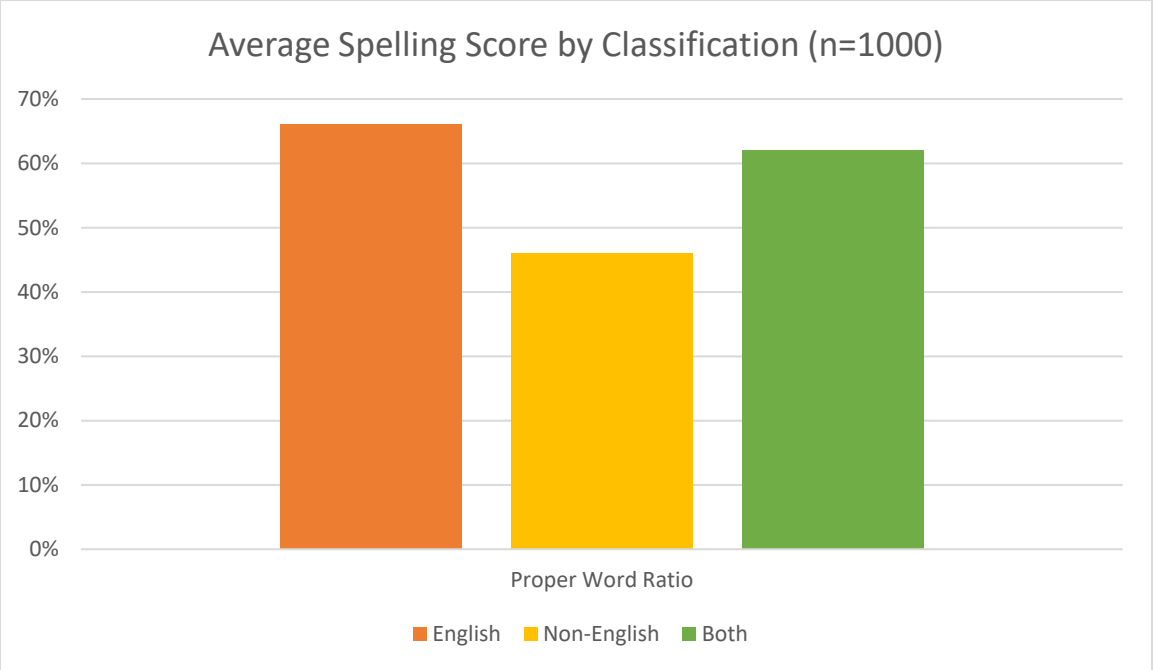
Our first notable observation is the difference in character length and word length between English and "non-English" tweets:



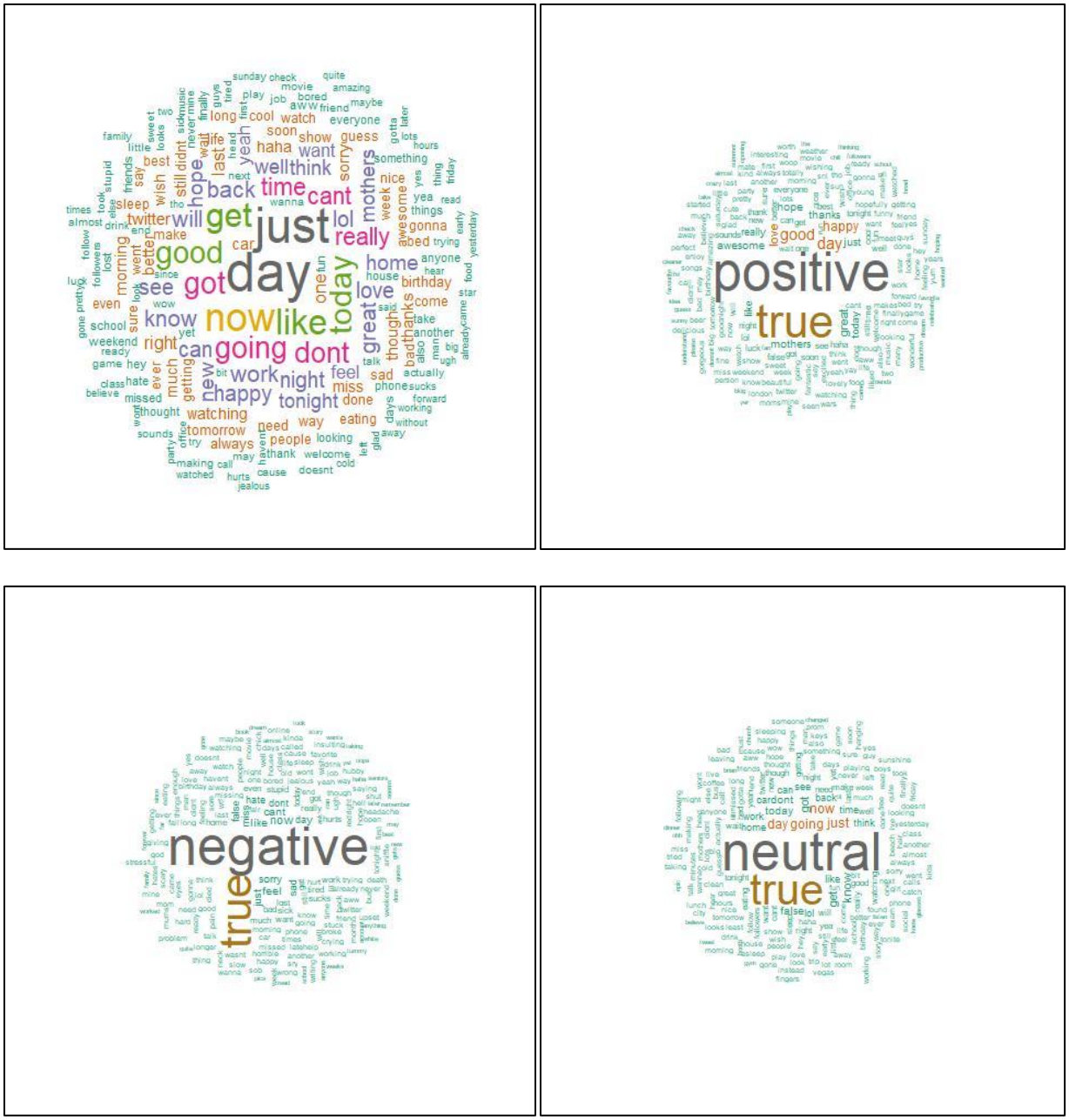




It seems intuitive that the fewer characters a tweet contains, the less likely a language identification algorithm is likely to pick the correct language, even an advanced one such as CDL. As a result, we can immediately extrapolate from our dataset that the fewer characters a tweet has, the less likely it is to both (a) be classified and parsed in its correct language and (b) have its polarity correctly identified, if at all.

The next reasonable step to take is to identify the relationship between proper classification and the frequency of correctly occurring words in a tweet. This runs into several problems – primarily, how to identify is a word is “correct” in a forum where slang and misspelling is a very rampant practice. Nevertheless, provided the chance of any given tweet containing “slang” is about the same, we ought to be able to demonstrate a relationship with some R scripting that utilizes a URL pull from the SCOWL dictionary of English Words. We can define a function which returns the ratio of correctly spelled words to all words in a tweet to track this relationship:




It’s clear that punctuation, word count, and character length all factor into the quality of our ABT. Although these issues will have contingencies during the modelling process, we should devote some time to observing the difference in word frequency by each of the polarities in our dataset as well – we can do this by filtering out stop words and punctuation and creating a word cloud.




**From left to right: Word Cloud for All Tweets, Positive Tweets, Negative Tweets, Neutral Tweets-**

**generated using the wordcloud package in R.**

Our word clouds show a very interesting result for each of our polarities – the name of the polarity itself seems to be the most frequent word! Otherwise, results seem relatively intuitive. Our positive cloud features words like hope, good, and awesome, while our negative cloud features words like sorry, sad, and hate. Our neutral cloud seems to not exhibit any clear

sentiment in any direction, which is largely in line with our expectations for this dataset. As a further modelling application, we can utilize word clouds from our training set to see which polarity a tweet is likely to represent given how much of the tweet can be associated with any given cloud.

Stop words, while they cannot provide too much insight for polarity analysis, can be used as a proxy for determining the quality of our data. Simply put, if our data contains a disproportionate amount of stop words per tweet, then we can suggest that the quality of our data is poor. Furthermore, it is possible that tweets with more stop words are more likely to be identified as neutral.

| Polarity | Percentage of Tweet in Stop Words |
|---|---|
| **Positive** | 27.7% |
| **Neutral** | 28.8% |
| **Negative** | 31.01% |

An off-hand analysis of the ratio for each polarity doesn't seem to suggest significant differences in stop-word usage, meaning we can extract little insight from stop word descriptors as a modelling dimension. Finally, we can look at the correlation between all the different dimensions we've attached onto our ABT as part of the exploration process:

| | Spelling Score | Stop Word Ratio | Tweet length | Word Count |
|---|---|---|---|---|
| Spelling Score | 1 | 0.6628 | 0.2773 | 0.325 |
| Stop Word Ratio | | 1 | 0.2795 | 0.3254 |
| Tweet Length | | | 1 | 0.95 |
| Word Count | | | | 1 |

Several notable observations emerge from this analysis:

1. Word Count and Tweet Length are highly correlated – this is obvious, and not particularly insightful to us, as they're essentially built from the same data.
2. Stop Word Ratio is moderately positively correlated with the spelling score – we can interpret this to mean that tweets with better spelling and grammar are easier to handle for any suite of NLP modules because it's easier to identify stop words when they're spelled correctly – as an added benefit, the fact that the rest of the tweet is more likely to be grammatically correct means that it ought to be easier to classify the language that it's in, as well as the polarity of the tweet itself.

**Challenges thus far:**

*Note: This section is to be modified dynamically as revisions to the EDA and data quality process persist.*

Overall, we've determined that about 80% of any given sample of data from our ABT is readily usable, suggesting moderate to good quality of data. The quality is co-opted by several problems, including the presence of undetected sarcasm, the presence of URLs and other non-word characters, the presence of stop words which generate noise in our future sentiment analysis, and the presence of shorter tweets which make it difficult to identify the language as well as the polarity.

Despite these shortcomings, we were able to generate a visualization of the relationship between certain grammatical dimensions that were contrived using outside R libraries and procure a word cloud for each type of sentiment that was able to provide insight into the type of words we can expect to associate with each polarity in our modelling.

.

## References

CloudFlare (N.D.). *What Is A Social Media Bot?* Retrieved January 25th, 2021 from https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/

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/

Lubin, Gus (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

Mousa, Mouhammad (July 21, 2016). *Sentiment Analysis: Is Sarcasm Positive, Neutral, or Negative?* CrowdAnalyzer, Retrieved January 25th, 2021 from https://www.crowdanalyzer.com/blog/sentiment-analysis

ParallelDots (N.D.), *SmartReader Analysis API*, accessed 1/25/2021 from https://smartreader.paralleldots.com/analysis

Pew Research Center (June 12, 2019) *Social Media Fact Sheet* Pew Research Center, Retrieved January 25, 2021 from https://www.pewresearch.org/internet/fact-sheet/social-media/

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