Capstone Project 2: Milestone Report
Spring Board Data Science Career Track
~ By Zach Nguyen

Data Story

Business Problem: Nowadays, companies juggle with quite a few product brands. Despite their maintenance cost, product brands are strong value generation engines for many established companies. For this reason, companies are constantly looking out across Blogs, Forums, and other Social media platforms, etc. to check the sentiment for their various products and competitor products to learn how their brand resonates in the market. This kind of analysis helps them as part of their post-launch market research to determine the effectiveness of the brand and the product they have released into the market.

Client: Even though this problem is significant in the medical industry, it can be very similar in many other consumer-driven industries like food, clothing retail, e-commerce etc .. With a good algorithm, these companies can flag good and bad products for review and generate real-time metrics on how their products are doing, which products are more salient and how well.

Data source: The data found is from a hackathon hosted by Analytics Vidhya which is already expired. Thus, we cannot generate the score report. However, we can track our performance with cross validation scores.
https://github.com/rajat5ranjan/AV-Innoplexus-Online-Hiring-Hackathon-Sentiment-Analysis/?fbclid=IwAR0ITaYyMNQPnn4QpoYJ6r-xPKYHPL0E-kwmzdn2_gcwGQDxuAtTh6MgYas

Methodology:

- Use text analysis techniques to explore the data and address any data imbalance (with packages such as Pandas, Seaborn, nltk, gensim or native arguments in algorithm).
- Build and explore various ML models to improve text classification (sklearn, keras-tensorflow).
- Devise a business plan and recommendation for the use case of the algorithm.

Deliverables: Slide Deck, Report, Code

Data Wrangling and Preprocessing

After defining and using a simple function to check random reviews of different sentiments, a few aspects were discovered that could hinder the quality and integrity of the data:

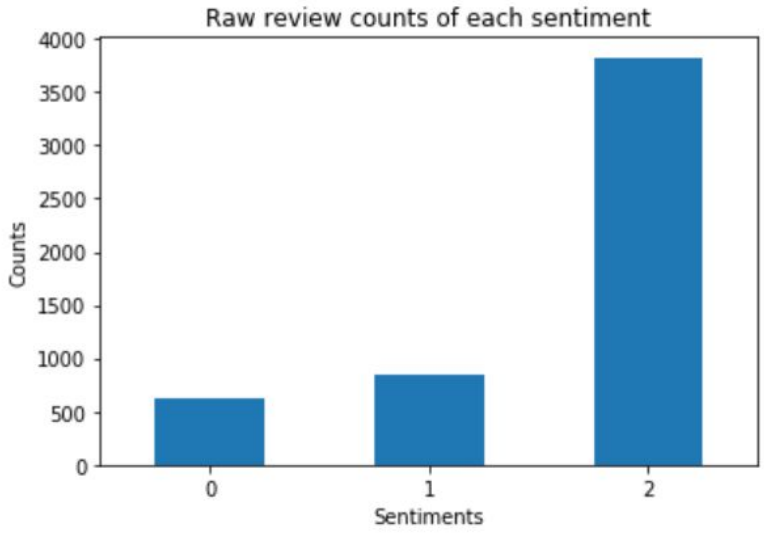
- None of the reviews are missing, so removal or imputation is unnecessary.
- The sentiments have three classes (0: positive, 1:negative, 2:neutral), thus multi-label algorithms are a must.
- The reviews include a high data imbalance, which could be rectified with resampling techniques to including appropriate weights in our algorithm.
- Many reviews and are confusing to classify, even for humans! This is due to the fact that a review could talk about other drugs (and therefore, could express different sentiment than the drug in question).
- The reviews include a lot of punctuations and numerics and occasional links/urls that could hinder Bag of Word analysis.
- The drug-name sometimes include hyphens and numerics, so caution must be taken to ensure the drug name remains in-tact.
- The reviews include a lot of contractions, which could create inaccurate lemma (example: "don't" is not the same as "do" when lemmatizing).

To resolve the above issues, the steps performed for data cleaning include:

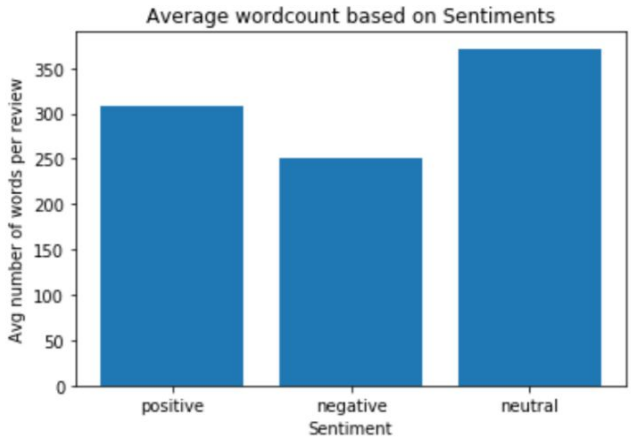
- Convert the text to lower-cased.
- Expand all contractions using a custom contraction mapping dictionary.
- Remove all links.
- Remove all punctuation except hyphens.
- Remove all numerics and hyphens which stand by themselves (between two spaces).
- Lemmatize each document with respect to its Part-of-speech using a lemmatizing function.

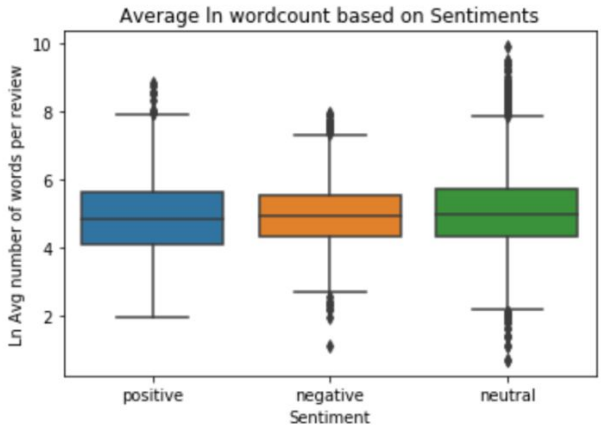
Exploratory Data Analysis

1. **Sentiment break-down:** Our sample displays high imbalance with a strong bias towards the neutral class. This tendency is surprising, since people don't usually write reviews for products they feel neutral about. Note, however, that the reviews here contain comments to other reviews as well. The reasonable next step is to look at the word count of sentiments.



2. **Average Word Count:** There looks to be a difference in the average word count of the different sentiments. Specifically, negative reviews tend to be more curt and neutral reviews more verbose. However, upon closer inspection using boxplot, we can see that the median wordcount and the interquartile ranges is quite close among all three sentiments. This suggests that neutral reviews tend to be more on the extreme.





A question arises as to whether the average word count is really statistically significant. Using the Kruskal-Wallis H test (The ANOVA one way cannot be used due to its assumptions of normality, which is clearly violated with our samples), we can be confident that they are with F-stat of 11.9 and a p-value of less than 0.05.

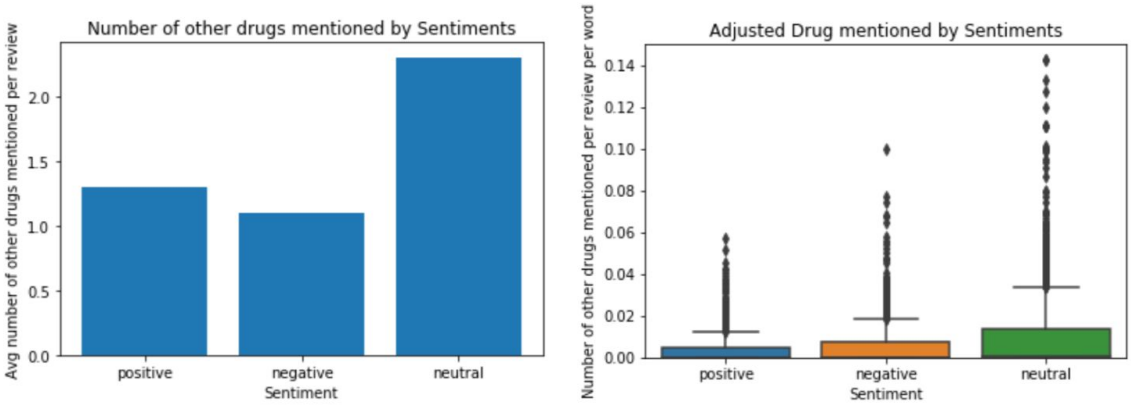
3. **Innate drug quality:** A question arises to whether each drug has innate quality that allows sentiment on them to be judged as negative/positive/neutral. Investigation shows that there are drugs with correspondingly 100% positive/negative/neutral reviews. Thus, it can be concluded that what type of drugs being reviewed can affect the sentiments. Here is an example of 10 drugs which only has neutral reviews:

```
# Sort the dictionary and display the items with highest neutral rates

sorted(neutral_rate.items(), key = lambda x: x[1], reverse = True)[0:10]
```

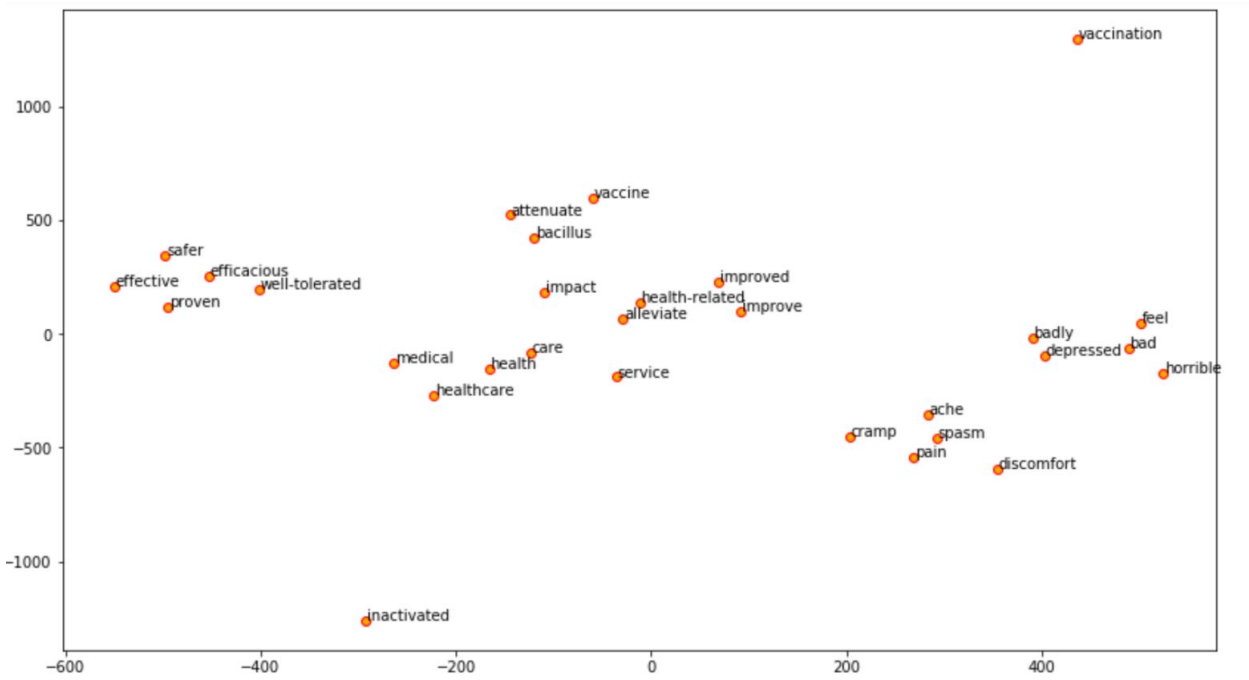
```
[('pan-retinal photocoagulation', 1.0),
 ('ipilimumab', 1.0),
 ('ixifi', 1.0),
 ('teriflunomide', 1.0),
 ('zykadia', 1.0),
 ('yervoy', 1.0),
 ('amjevita', 1.0),
 ('pemrolizumab', 1.0),
 ('tafinlar', 1.0),
 ('gilotrif', 1.0)]
```

4. **Number of drugs mentioned adjusted for length of review:** Another question arises to whether or not other drugs mentioned in the review contribute to its sentiments. An example is a review comment which lists out many different types of drugs as an objective (neutral) suggestion. Neutral reviews seem to have high drug mentions, even when adjusted for number of words in such review.



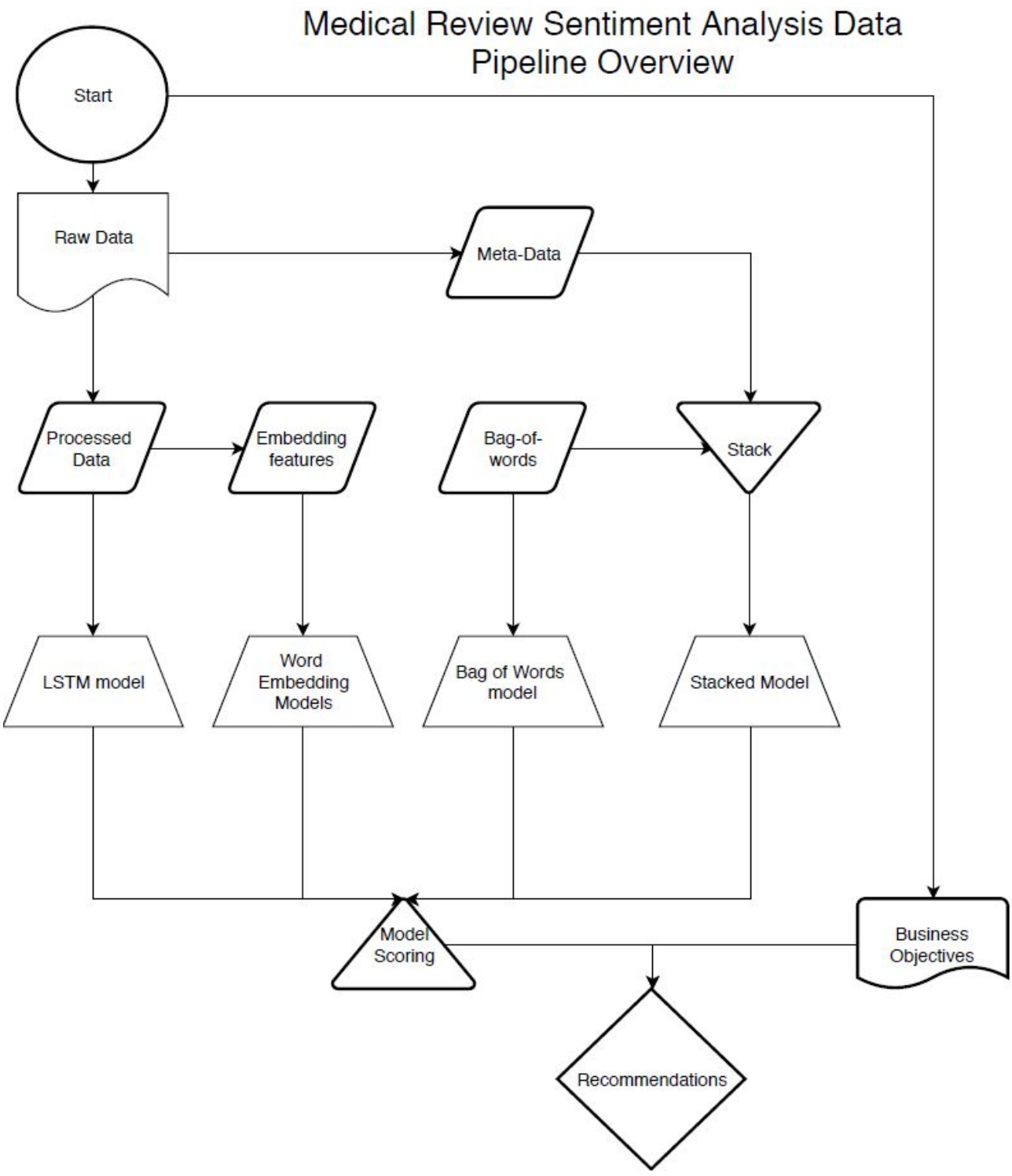
Once again, Kruskal-Wallis H test reinforces our belief with a p-value close to 0.

5. Word2Vec visualization: Due to the difficulty of this sentiment analysis task (it is difficult even for me, a human, to classify many of the sample reviews), we will want to know if our data is discriminative enough to provide value for the business. If not, the business is better off pursuing other opportunities. A few ways to visualize the discriminativeness of text data is to use WordCloud, LDA topic modelling and Word embeddings. This time, we will use TSNE dimensionality reduction technique to visualize word embeddings. To prevent clutter, I will use 2 words of each sentiment ('health', 'vaccine', 'effective', 'improve', 'bad', 'pain') and their 4 similar words to visualize the spread in two dimensions:



The result shows some distinctions. Positive words are on the middle and top left, whereas negative sentiments are in the bottom right. However, some neutral words are blurred with positive words, which could increase the possibility of misclassifications.

6. **Dataset conclusion and plan:** With the above analysis, it we were able to discover some discriminative features despite the foreseen difficulty in the classification task. The following plan is put forward as an attempt to build a valuable sentiment classification algorithm. In general, we will conduct simultaneously data processing and business alignment. On one hand, with data processing, we will extract meta-data (which contains some discriminative information), bag of words features and word embeddings features to run cross validated grid-search on four families of algorithms (BoW models, Meta-data Stacked models, Word Embedding Models and Neural Network). The goal is to find the strongest all-around classification architecture measured by F1 Macro Score. On the other hand, we will consult with business personnel to analyze the cost-benefit of ML integration and build a Pay-Off matrix to translate the result of the algorithm into business values. In the end, we will tweek the best algorithm using custom architecture and make recommendations which can align with business objectives. The Overview of the project is outlined in the diagram below:


Medical Review Sentiment Analysis Data Pipeline Overview
Start
Raw Data
Meta-Data
Processed Data
Embedding features
Bag-of-words
Stack
LSTM model
Word Embedding Models
Bag of Words model
Stacked Model
Model Scoring
Business Objectives
Recommendations