Understanding Posts on Twitter: Plant-Based/Vegan Meat

Abstract

The goal of this project is to utilize text data and transform it using NLTK tools to produce interpretable results that capture a deeper analysis on the topic of plant-based/vegan meat keywords from twitter data. As the market for plant-based/vegan meat products is on the rise, the question on how the public views these products are highlighted. Brand deals with fast food franchises are being marketed and scientific resources present evidence on the increasing benefits plant-based/vegan meat does for the environment. Topic modeling was used to distinguish key topics along with its corresponding words while sentiment analysis using Vader brought more insight on how topics change in regards to sentiment over time. The usage of ScatterText was to further gain understanding on which distinguished terms were associated with positive vs negative sentiment.

Design

The project utilizes snscrape python library to scrape twitter data based on keywords: plant-based meat and vegan meat from 2020 to recent. With personal experience on eating vegan meat products from Beyond Meat or Impossible Foods, I was interested in learning how others view these products as well. Aside from my personal interests, the project can be used to formulate marketing strategies from content that is not filtered.

Data

The dataset comprises 33,580 entries with 5 features (datetime, url, content, username, like_count) each representing a post from a user from 2020-recent (October 2022). After data cleaning, the final dataset contains 28,873 posts with 12 features. The added features include: cleaned text, sentiment score, sentiment label, dominant topic, topic terms and topic number.

Algorithms

Text Preprocessing

Due to verbose language in tweet posts, text preprocessing was vital to provide optimal context of each source. Punctuation, capitalization, removing emojis, url, hashtag, and tags were built in through a function that would go over each row of the data to process the clean content. A customized stop word list was an important aspect to capture all words that would not provide distinguished information. The Metis stopword list was used in coherence to the provided stopword list from NLTK.

Models + Defining Selected Model

NLTK was used to help compare different models like LSA and NMF for topic modeling. CountVectorizer and TfidfVectorizer were tools that were also compared against the two models. NMF w/ TfidfVectorizer was selected due to topics being more interpretable. The parameter was tuned by using source code (Derek Greene github) where genism was used. 11 topics showed to have the highest coherence score and among them, 15 words for each topic were selected. The 11 topics are the following: Healthy Diet, Fake Food, Brand Deal, Processed Food, Products, Alternative Option, Healthy Eating, Diet Options, Taste, People's Choice, and Protein Source.

Sentiment Analysis

To get a better understanding on how the public views plant-based/vegan meat, VADER was used to calculate compound score across each document. Overall, all topics generate a neutral to positive sentiment as well as throughout all years from 2020-2022. When investigating negative tweets on certain topics, taste and how Burger King's Impossible Whopper being cooked on the same grill as meat menus propagate general negative sentiment.

ScatterText

The interactive webpage showcases frequency of word count throughout all posts from twitter. Positive vs Negative category shows the expected word list that represents each category. There are times when certain negative posts seem to underlie support for vegan meat, but due to words that correlate in a negative manner the sentiment captures negative sentiment.

Tools

- Numpy and Pandas for data acquisition and manipulation
- Scikit-learn (NLTK) to model
- Seaborn, Matplotlib, WordCloud for visualizations

Communication

Will host on my github.