NLP Final Project:
News via Sentiment Analysis & Taxonomy-Guided Semantic Similarity

Oluwademilade (Demilade) Adeboye, Zhikun (Devin) Chen, Michael Latimer, Junkai (Frankie) Lin

Background & Use Case

Focus Area

Investors and Quantitative Traders relies on the news to make decisions about where to invest and what strategies will yield higher returns.

Process

Algorithm classifies topics into 11 industries, provides overall sentiment for the respective industry, and summarizes most similar articles to search query.

Output 1

Real time sentiment on hundreds of topics with corresponding ETFs for quantitative traders to develop strategies of trading based on NLP sentiment analysis.

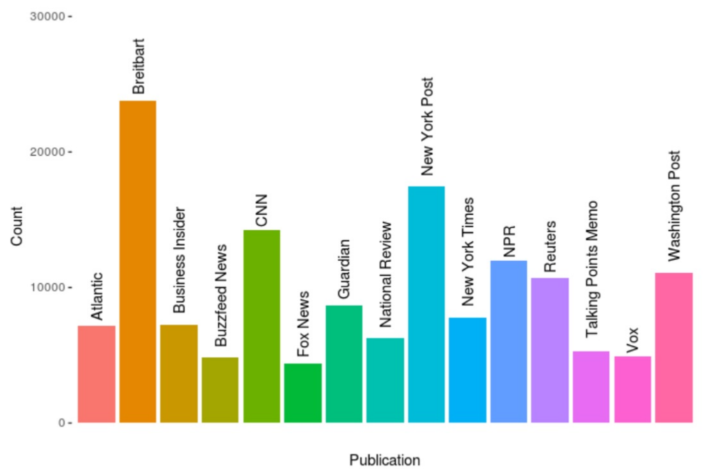
Output 2

Investors can also search for the news topics and industries for the whole content of the news and their sentiment score to get more information on the industries.

Data Source & Procurement

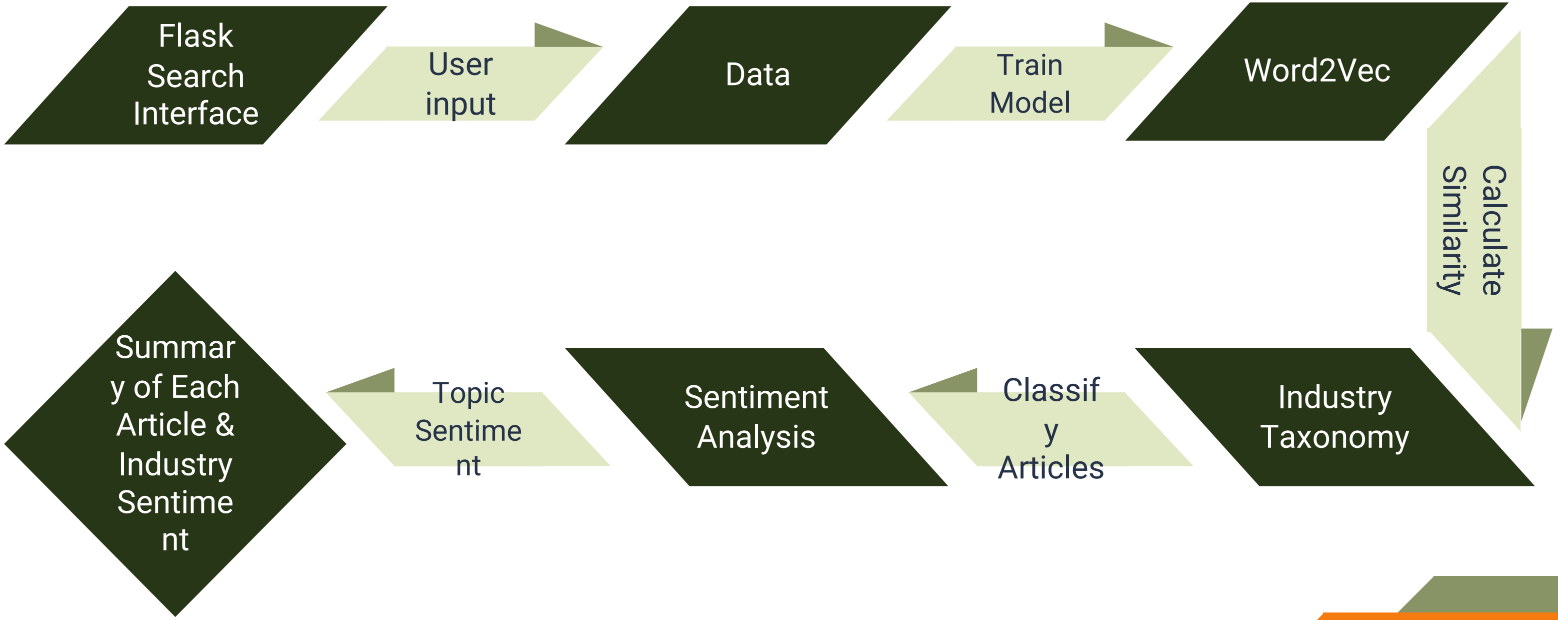
All the News Data Set (669MB)

- 143,000 Articles across 15 publications
- Reuters data includes over 10,600 articles from 2015 to 2017. Articles span various industries and topics
- Structured Data - Columns include id, title, author, year, content
- Dataset includes publications such as CNN, NYT, BreitBart and Reuters



https://www.kaggle.com/datasets/snapcrack/all-the-news?resource=download

Models & Methods



Flask Search Interface

User input

Data

Train Model

Word2Vec

Calculate Similarity

Industry Taxonomy

Classify Articles

Sentiment Analysis

Topic Sentiment

Summary of Each Article & Industry Sentiment

Technology Specifications

Taxonomy

Classifies articles into 11 industries and 40 plus sectors.

Taxonomy key words derived from domain knowledge and SEO websites.

Keywords divided into subsections for each industry (i.e. consumer staples & soft drinks)

Sentiment Analysis

SentimentAnalyzer from NLTK package used to determine sentiment.

Sentiment based on compound score from SentimentAnalyzer.

NRCLexicon predicts the sentiments and emotion of a given text

Word2Vec

Similarity calculated using trained Word2Vec model.

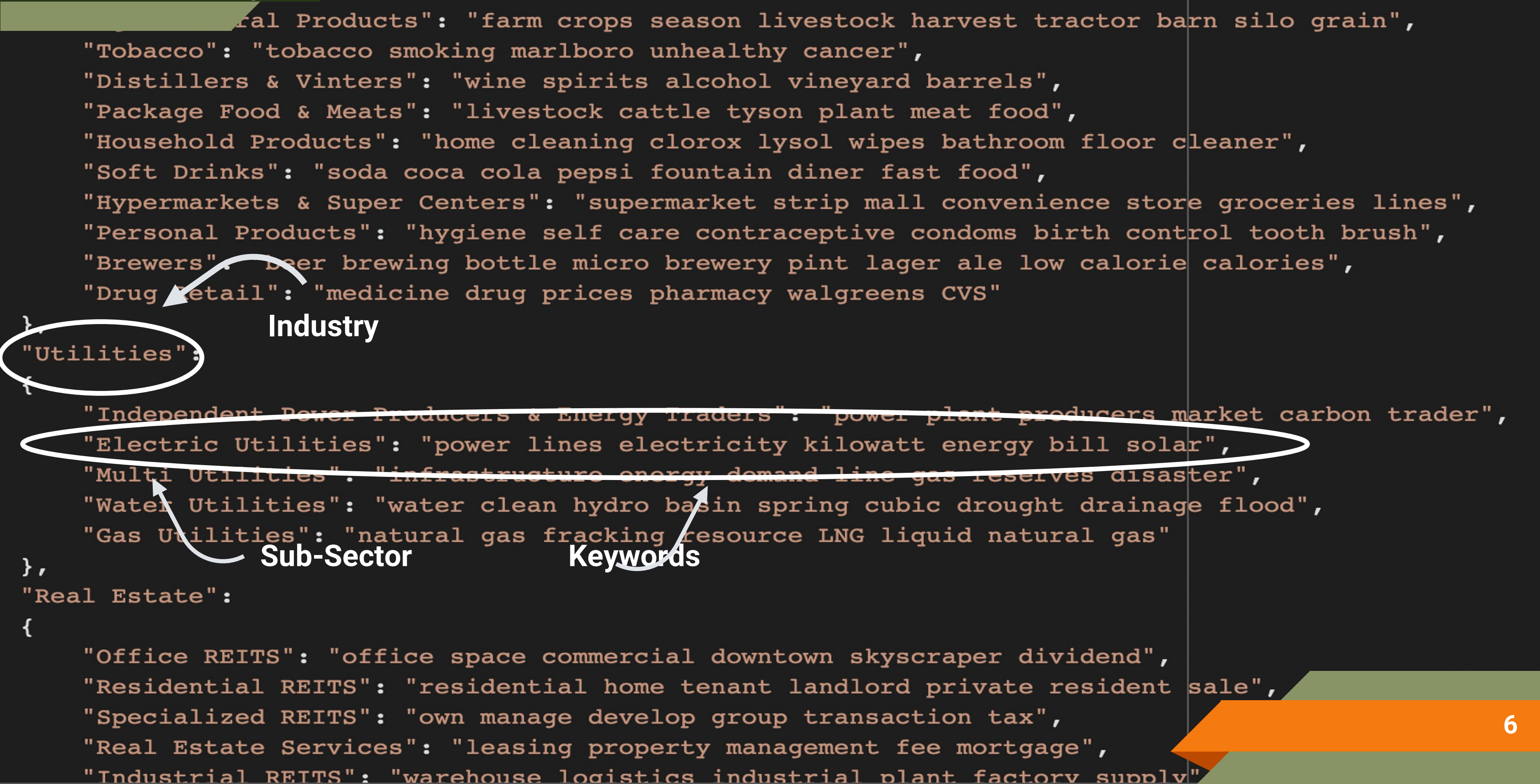
Word2Vec model trained using Gensim package.

Summarizatio n

Summarization using FastT5 model

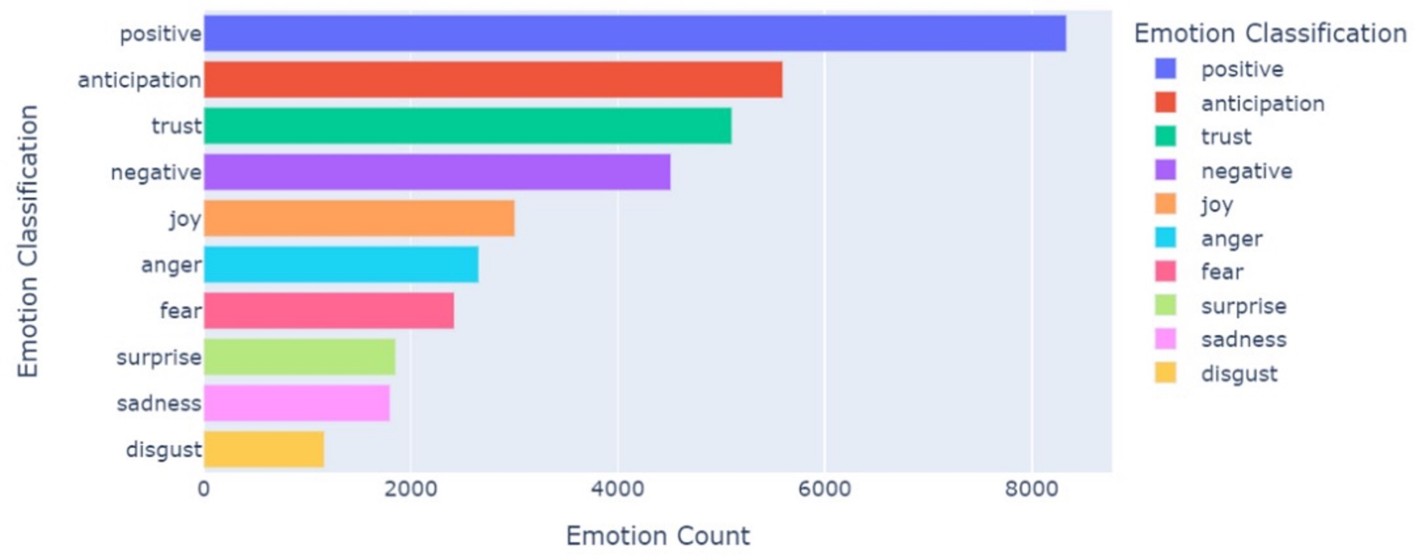
FastT5 model summarizes the content of article which condense a range of information, giving readers an aggregation of the most important parts of what they're about to read

Taxonomy

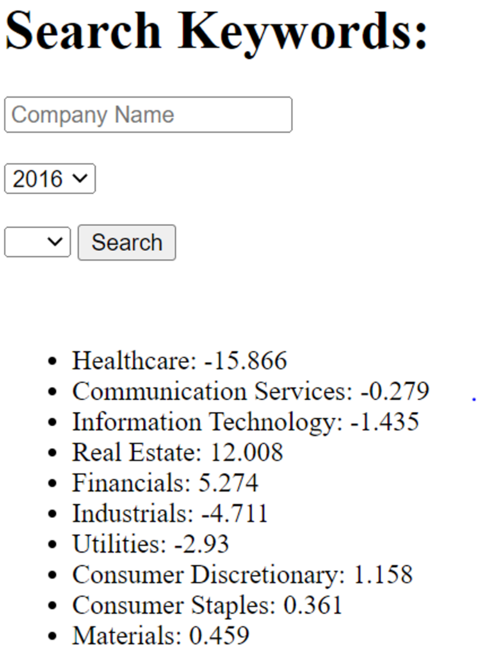


```
ral Products": "farm crops season livestock harvest tractor barn silo grain",
    "Tobacco": "tobacco smoking marlboro unhealthy cancer",
    "Distillers & Vinters": "wine spirits alcohol vineyard barrels",
    "Package Food & Meats": "livestock cattle tyson plant meat food",
    "Household Products": "home cleaning clorox lysol wipes bathroom floor cleaner",
    "Soft Drinks": "soda coca cola pepsi fountain diner fast food",
    "Hypermarkets & Super Centers": "supermarket strip mall convenience store groceries lines",
    "Personal Products": "hygiene self care contraceptive condoms birth control tooth brush",
    "Brewers": "beer brewing bottle micro brewery pint lager ale low calorie calories",
    "Drug Retail": "medicine drug prices pharmacy walgreens CVS"
},
"Utilities":
{
    "Independent Power Producers & Energy Traders": "power plant producers market carbon trader",
    "Electric Utilities": "power lines electricity kilowatt energy bill solar",
    "Multi Utilities": "infrastructure energy demand line gas reserves disaster",
    "Water Utilities": "water clean hydro basin spring cubic drought drainage flood",
    "Gas Utilities": "natural gas fracking resource LNG liquid natural gas"
},
"Real Estate":
{
    "Office REITS": "office space commercial downtown skyscraper dividend",
    "Residential REITS": "residential home tenant landlord private resident sale",
    "Specialized REITS": "own manage develop group transaction tax",
    "Real Estate Services": "leasing property management fee mortgage",
    "Industrial REITS": "warehouse logistics industrial plant factory supply"
```

Sentiment Analysis



Sentiment Analysis

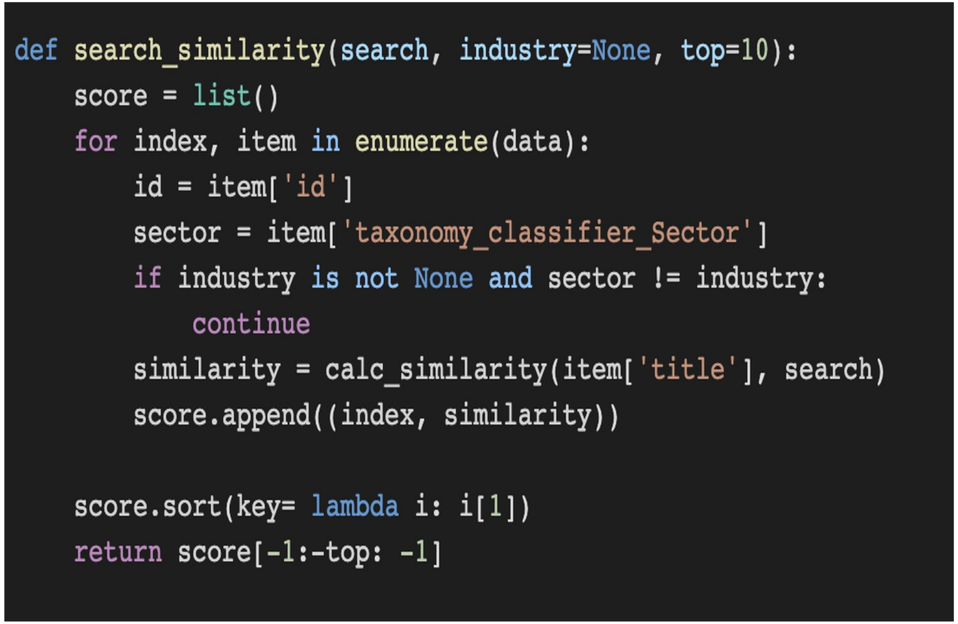


- Healthcare: -15.866
- Communication Services: -0.279
- Information Technology: -1.435
- Real Estate: 12.008
- Financials: 5.274
- Industrials: -4.711
- Utilities: -2.93
- Consumer Discretionary: 1.158
- Consumer Staples: 0.361
- Materials: 0.459

Package	NLTK
Function	Sentiment Intensity Analyzer (Polarity)
Score	Compound

Word2Vec & Similarity Scoring

- Similarity calculated against Google News Word2Vec pretrained model
- Function (right) calculates similarity based on keyword & industry searched and returns top 10
- Selected articles passed onto Flask and displayed along with compounded sentiment score



```
def search_similarity(search, industry=None, top=10):
    score = list()
    for index, item in enumerate(data):
        id = item['id']
        sector = item['taxonomy_classifier_Sector']
        if industry is not None and sector != industry:
            continue
        similarity = calc_similarity(item['title'], search)
        score.append((index, similarity))

    score.sort(key= lambda i: i[1])
    return score[-1:-top: -1]
```

Product Demonstration

Bear with us :)

Evaluation Criteria

70%

Accuracy (precision)

We tested our class taxonomy based on 100 article test set with over 5000 training set and scored 70% accuracy on the topic classification.

High accuracy topic classification helps derive the more precise sentiment score.

2 Sec

Avg Search Time

We achieved an average word2vec search time of 2 seconds.

Search time slowed by Word2Vec similarity calculation

Future Considerations

- Current summarization is computationally intensive and time consuming, future work should explore faster methods
- Our current work was limited by the articles available, and we anticipate scaling this product to work with larger sets of articles
- Utilize News API to search for relevant, recent articles before running sentiment analysis and similarity scoring
- Limit the scope of the project to focus on quantitative trading, allowing users to track current sentiment and update models from unbiased sources