Better Dwelling Week1 Milestone - Project Plan

Team members: Pandramish Naga Sirisha, Amy Lam, Jonathan Chan, Aaron Tian
Date: May 12, 2020

1. Description

In this project “Does News Report or Manipulate Financial Markets?”, we want to analyze the trends in the six economic indicators of our interest i.e., mortgage rates, interest rates, house prices, employment, GDP, TSX index value with respect to the sentiment of the economic news articles. Our primary interest lies in visualizing the effect of the sentiment of the economic news pertaining to real estate data on the shift of the economic indicators as a result of the publishing of the articles before or after the indicators show the change.

We aim to use a sentiment analyzer and analyze the correlation of these six economic indicators with respect to the sentiment of the news articles for each of the indicators that is listed above. In the first step, we use a sentiment analyzer. For this model, the dataset is extracted from scraping the web and identifying news articles that are relevant to the indicators. Then we build a model to predict the sentiment of these news articles. The predictions would be the sentiment of the article on an ordinal scale. As we now have the economic indicators as published by the government resources and the predicted sentiment scores, we measure the correlation of the change in the economic indicators with respect to the sentiment of the articles published segregated by source on a temporal scale.

2. Datasets

In collecting articles for analysis, the project will initially focus on two Canadian news sources: CBC.ca and Bloomberg.ca. Both outlets were targeted based on the intent to gauge sentiment of prominent Canadian news outlets while also collecting articles likely to be heavily related to the financial indicators of interest. Availability of publicly accessible articles and supervisor recommendations also contributed to selecting these news outlets. Based on initial findings, other news outlets to consider include: Global News, CTV news, National Post, The Globe and Mail.

Articles will be scraped and stored in json format (dictionary of strings and datetime items). In addition to the article content, several key pieces of metadata will be required for each article:

Author name, publishing date, URL, subtitle, and title. Data will be scraped and stored in the project github repo.

3. Expected deliverables

As stated in the capstone project outlined by Better Dwelling, the core deliverables should be (1) a project pipeline for the economic news sentiment scores and the associated interactive visualization, and (2) a report outlining the processes and findings.

For the project pipeline end product, core deliverables include: (a) a backend corpus that includes the news articles we scraped and stored in JSON format that are consistent with Better Dwelling's existing corpus, (b) a frontend interface website to index the corpus with search keywords and associated visualizations of correlation analysis, (c) a baseline sentiment analysis model with good accuracy and (d) codes to handle web scraping, data cleaning, data storage, and visualization.

If time allows, additional deliverables can include: (e) a set of aspect-based sentiment classifiers each fine tuned for a specific macro-economic indicator; (f) Inclusion of more macro-economic indicators for correlation analysis and visualization; (g) expansion of dataset to include more Canadian news sources.

4. Methods

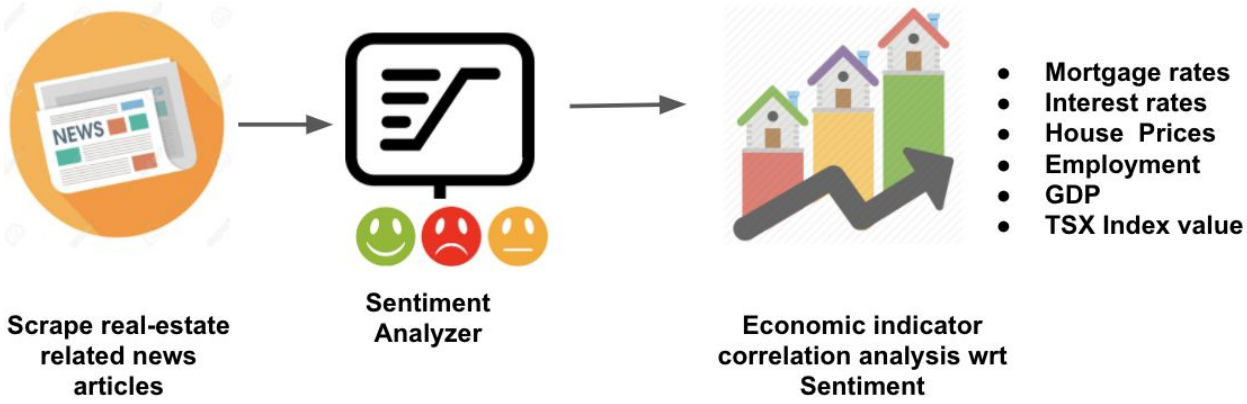

Figure 1 - Major project components

Data Collection and Wrangling

The project team will use Python libraries such as urllib [2], Requests [3], and BeautifulSoup [4] to collect different pools of business articles, each targeting a specific economic indicator such as mortgage rates, interest rates, housing prices, etc. Data wrangling will be performed using data structures such as Pandas DataFrames [5]. News articles collected in the initial phase will be manually annotated for sentiment polarity and used for model fine tuning. The scale and granularity of the sentiment polarity score will be decided later depending on the sentiment analyzer we use. The sanitised and annotated data will be saved in JSON format.

Computing Infrastructure

The project team will conduct web scraping, data wrangling, data preprocessing, and initial model building on personal computers. Group members will run and test deep learning models on Google Colab.

Sentiment Analyzer

The project team will employ a Bidirectional Encoder Representations from Transformers (BERT [6]) based sentiment analyzer such as flair [7], that could output sentiment scores of business articles on an ordinal scale, in which case packages such as Pytorch [8] and Transformers [9] will be used. The BERT was pretrained on a massive amount of generalistic text and in this case will be fine tuned on a dataset of business articles as well as a benchmark dataset that was created for sentiment analysis purposes. Considering BERT based methods are resource-heavy and the team is using laptops to work on this project, more traditional sentiment analysis tools such as TextBlob [10] or VaderSentiment [11] can be used for the sentiment analyzer. We will make it easy for the actual users of this pipeline to substitute the sentiment analyzer with a state of the art model if they have more computing power. Since the analyzer will generate sentiment score on an ordinal scale, ordinal classification methods such as SVM ranking might be used here. Individual articles could include content that relates to more than a single economic indicator. In this case, aspect-based sentiment analysis methods will be considered as well.

Weights for Multiple Sources

Correlation-generated weightings

By default, the system will generate weights for each news source regarding an economic indicator based on the historical correlation of trends between the sentiment score of that news source with the particular economic indicator. For example, assuming there are two sources for an economic indicator, if source one has a 0.7 correlation and source 2 has 0.35 correlation, then we could weigh source one by 0.66 and source 2 by 0.33. Volume of data points regarding a particular economic indicator from different sources will be another factor to consider when determining the weights.

Rough example calculation:
https://docs.google.com/spreadsheets/d/1qJJVXlRV850GvHg-lEz_EQoZ8co_maBQ-W8Ic9K2lIA/edit#gid=0

User-defined weightings

When users are using our visualization tool, they will be able to replace the default correlation-generated weights with weights of their choice. Interface users can input weights for each source to modify a particular source's importance when calculating final sentiment score.

Code Review Checklist

This checklist will be used for the team to review the code written by team members. This is used to maintain coding standards and nomenclature throughout the project. We create issues in the github repository for the code reviews and mark as resolved when the issue is resolved.

- ✓ Does the code do what it is intended to do?
- ✓ Is the code modular?
- ✓ Is there documentation for all the parts of the code?
- ✓ Are there test cases written in the code?
- ✓ Does the code cover edge cases?
- ✓ Is the code reusable? If yes, how? If not, how to make it more reusable?
- ✓ Is the code modular? If not, provide ways to make the code more modular
- ✓ Does the code include citations in case it has been referenced from other sources?
- ✓ Is the same code duplicated more than twice?
- ✓ Are the variable names used in the program appropriate to convey intent?

Visualization

Figure 2 - Initial outline of visualization components

Final visualization will be created using Javascript and D3.js [12]. Visualization interface should allow a user to provide the following inputs:

- Sentiment score - selection of news articles with a sentiment score within a particular range
- Financial indicator - selection from 6 financial indicator options to display on time series data on
- Time period - selection of time period to display data from, segmented based on time intervals used in recording the financial indicators (daily, monthly, quarterly)
- Weighting method - selection of method of weighting method when calculating final sentiment score. Users can either select their own weights for each source or use system weights which are calculated using the correlation value for each source relative to the selected financial indicator.

Final visualization will display values for the selected financial indicator along with the sentiment scores for the articles published within the selected timeframe. Visualized data should change dynamically as the user inputs are modified, with time series intervals corresponding to how

data was collected for the selected financial indicator. Weight values for each news source should be displayed in the final visualization.

5. Schedule

	Task	Due date	Mini project manager
Week1 (May 4 - May10)	Project Plan	May10	Sirisha, Amy
Week2 (May 11- May 17)	Corpus Collection Annotation; initial experimental time for model(on SOTA sentiment classifiers)	May 17	Amy, Aaron
Week3 (May 18- May 24)	Final corpus collection and storage	May 24	Aaron, Jon
Week4 (May 25 - May 31)	Sentiment Analysis Model fine-tuning; Correlation Analysis of variables and macro-econ indicators	May 31	Sirisha, Aaron
Week5 (Jun 1 - Jun 7)	Visualization/ Interface	Jun 7	Jon, Amy
Week6 (Jun 8 - Jun 14)	Code cleaning, video presentation preparation	Jun 14	Sirisha, Aaron
Week7 (Jun 15 - Jun 21)	Writing final report and preparing presentation	Jun 21	Sirisha, Jon
Week 8 (Jun 22 - Jun 28)	End product deliveries	Data Product due on Tuesday June 23, 2020 18:00 -- To mentor (ungraded); Sunday June 28, 2020 18:00 -- To partner and mentor (graded)	All team members

		Video presentation due Thursday June 25, 2020 18:00	

References:

[1] https://github.ubc.ca/MDS-CL-2019-20/Capstones/blob/master/Capstone%20Projects/Does_News_Report_or_Manipulate_Financial_Markets_.md

[2] https://docs.python.org/3/library/urllib.html

[3] https://requests.readthedocs.io/en/master/

[4] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[5] https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html

[6] https://arxiv.org/pdf/1810.04805.pdf

[7] https://github.com/flairNLP/flair

[8] https://pytorch.org/

[9] https://huggingface.co/transformers/

[10] https://textblob.readthedocs.io/en/dev/

[11] https://github.com/cjhutto/vaderSentiment

[12] https://d3js.org/

[13] https://medium.com/palantir/code-review-best-practices-19e02780015f

[14] https://www.evoketechnologies.com/blog/code-review-checklist-perform-effective-code-reviews/

[15]https://gallery.azure.ai/Experiment/Predictive-Experiment-Mini-Twitter-sentiment-analysis-2

[16]https://www.iconfinder.com/icons/2598452/housing_market_graph_housing_market_stats_infographic_property_price_trends_real_estate_statistics_icon

[17]https://www.onlinewebfonts.com/icon/395879