Dylan Saez
May 02 2022

Measuring Sentiment of Recession-Based Keywords Using News Articles and NLP - REPORT

Introduction. An overview of the project and an outline of the report.

Economists, data scientists, and even non-technical individuals want to understand market behavior leading up to a recession or economic bubble. The Financial Crisis of 2007 – 2009 refers to a period of economic turmoil within the United States. A housing market bubble was the main character in the financial bust. Many banks around the world took on risky, unsustainable mortgages that banks gave out on a sub-prime basis. These banks wanted to fill their portfolio with credit and money owed back in an economy they thought would be able to re-pay those debts. Unfortunately, once the housing market was exposed to be insufficiently liquid, banks incurred a run-on and the housing market turn upside down from borrowers losing their homes to investors realizing their cash was never going to turn into profit.

In the present day, the post COVID-19 world looks a bit different but still runs very tight to how the GFC turned the markets. Price increases from global supply disruptions due to COVID-19 and the Russian aggression in Ukraine, oil prices increasing due to shortages and consumer demand mismatch, and inflation due to an economy pumped with rescue money only to be insufficient to the increased price of living.

Recovery from both events are slow – the metaphor *rockets and feathers* applies to both. Prices and economic downturn tends to escalate very quickly like a rocket launching, but the recovery is slow and non-linear, like a feather falling.

This project aims to take the top words from each batch of event news articles (GFC and post COVID-19) and measure their sentiment from a year before the *main event* to a year after the *main event*. The associations outlined from this project will help readers to conclude how terms are perceived (their sentiment) when reported upon in news articles and what that typically indicates of the economic current state.

The outline of the report is as follows:
1. Description of the data set
2. Description of the NLP model and what algorithms were used
3. Experimental set: how the data was used
4. Hyper-parameter specifics
5. Results
6. Summary and Conclusions
7. References

The outline of the project is as follows:
1. Collecting raw, labelled data
2. Pre-processing
3. Numerical Encoding of Text

4. ML Algorithms
5. Hyper-tuning and Training of ML models
6. Prediction

Description of the data set.

The data set was acquired through a uniform method presented in the literature. In the research paper, *Measuring Social Unrest Using Media Reports* by Barret et al. (2021) they acquired the news articles for their data set through Factiva, a Dow Jones news aggregator. In the paper, the authors restrict the database using specific keywords to identify the subjects the news articles needed to be categorized by, specific news organizations, and a region (for this project, re=United States).

The news organizations follow as such:

- Los Angeles Times
- Boston Globe
- New York Times
- Wall Street Journal
- Chicago Tribune
- Washington Post
- ABC Network
- CBS Network
- NBC Network
- USA Today
- Miami Herald
- San Francisco Chronicle
- The Dallas Morning News

Region = United States

Identifiers

- Domestic/Foreign Markets
- Mortgage Refinancing
- Mortgage Planning
- Market Research
- Small/Medium Businesses
- Asset-Backed Securities
- Economic Growth/Recession

These restrictions were applied to both batch acquisitions. For the GFC, date range: 01-01-2006 to 12-31-2010. For post COVID-19, date range: 01-01-2019 to 04-30-2022.

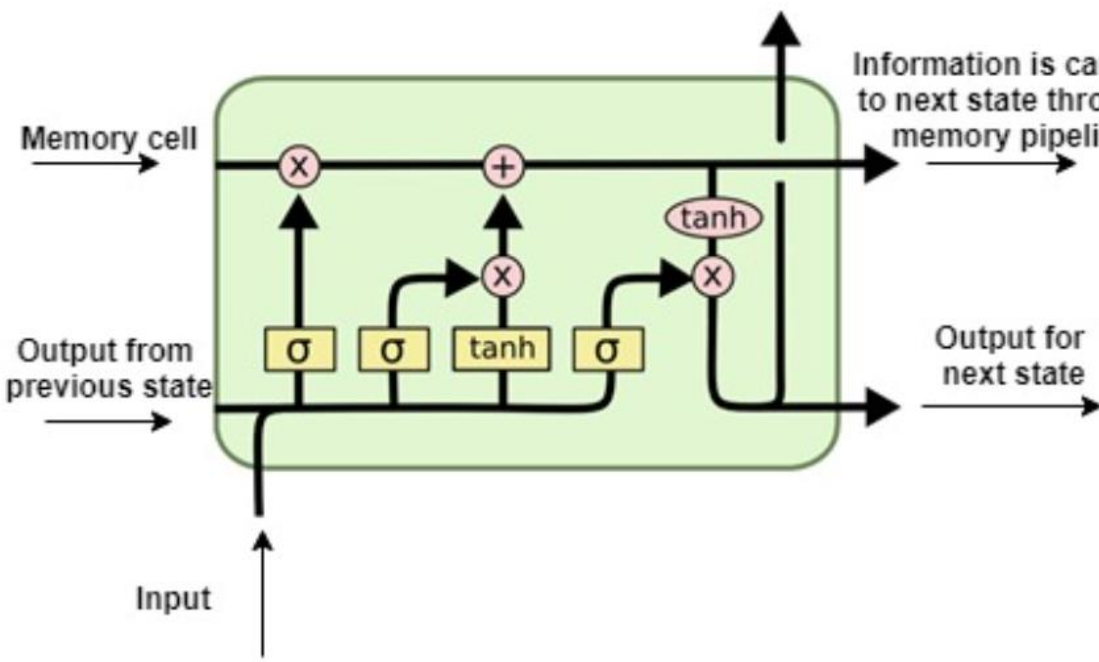
GFC # of articles: 298,589
Post COVID-19: 490,483

The GFC batch will be used as the training set. This means, the algorithms will be TRAINED upon this batch and then tested upon the batch of the post COVID-19 set to determine the sentiment of such terms and articles. The GFC batch was tagged with “positive” or “neutral” prior to inputting into the model.

Description of the NLP model and what kind of algorithm did you use. Provide some background information on the development of the algorithm and include necessary equations and figures.

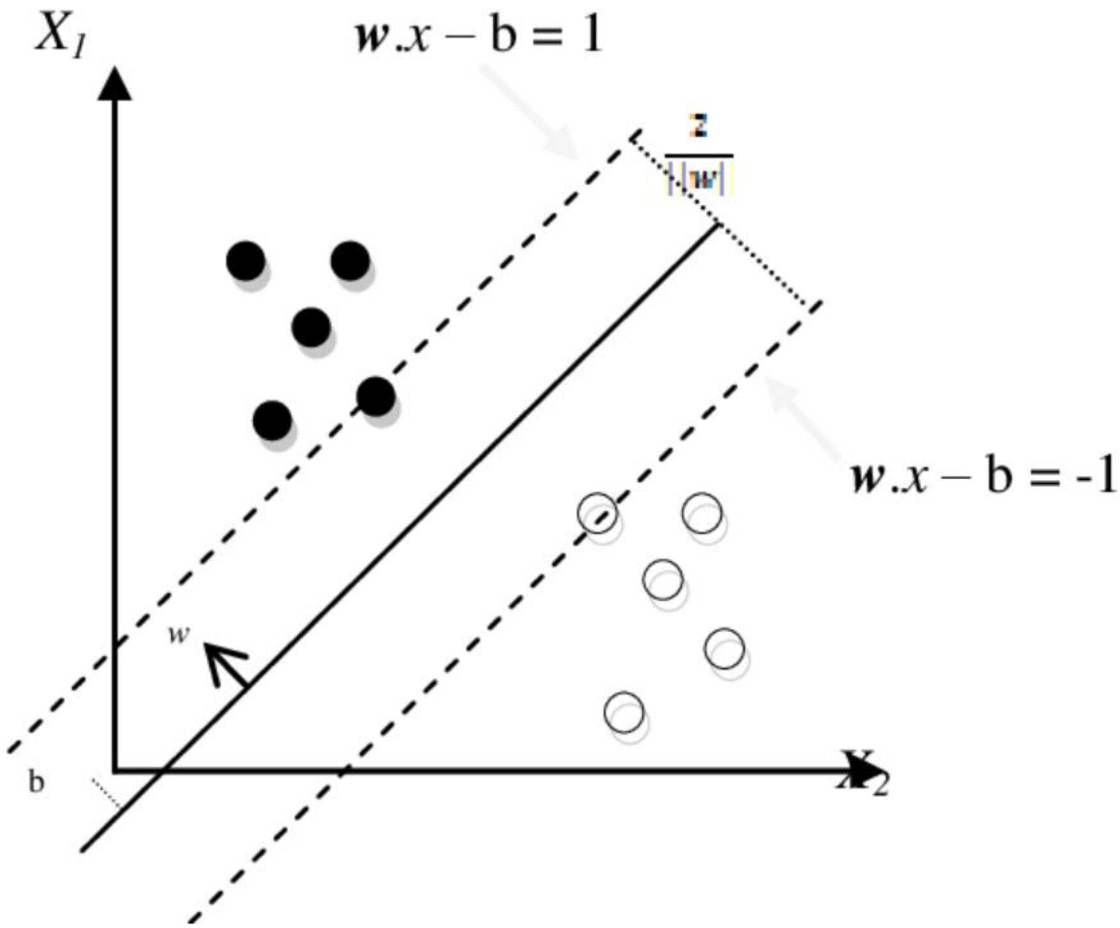
In this project, three models were used to test the post COVID-19 batch for sentiment analysis. LSTM, SVM, and Naïve Bayes.

Long Short Term Memory (LSTM) was designed to help with the problem of vanishing gradient of Recurrent Neural Networks (RNNs). In regards to RNN, vanishing gradient moves the gradient smaller and smaller until it has no real meaning. In other words, the marginal return to improving the model becomes useless.



Source: **https://colah.github.io/posts/2015-08-Understanding-LSTMs/**

Support-Vector Machines (SVM) is a supervised machine learning algorithm that is dynamic enough to be used for classification cases and regression analysis in natural language processing. In this project, classification (labeling of news articles) is the use-case. Performing classification with SVM is done by finding the hyper-plane that separates each class.



Source: https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1#:~:text=Sentiment%20Analysis%20is%20the%20NLP%20technique%20that%20performs%20on%20the,positive%2C%20negative%2C%20or%20neutral.

Naïve Bayes is very efficient at classifying large chunks of data. The algorithm mis based off of the Bayes probability theorem for unknown class prediction. Classes (or labels) are assigned within the naïve bayes algorithm as being the class with the highest probability.

For a hypothesis with two occurrences A and B, the MAP is

MAP (A)

= max (P (A | B))

= max (P (B | A) * P (A))/P (B)

= max (P (B | A) * P (A)

P (B) stands for probability of evidence. It's utilized to make the outcome more normal. It has no effect on the outcome if it is removed.

Source: https://www.analyticsvidhya.com/blog/2021/07/performing-sentiment-analysis-with-naive-bayes-classifier/

After splitting and vectorizing the text using tf-idf into numbers, the model using the training set and performs predictions on the test set features.

Before moving onto the *Experimental Setup*, let's define and explain tf-idf or Term Frequency – Inverse Document Frequency.

$$tf\,(t, d) = log\,(1 + freq\,(t, d))$$

$$idf\,(t, D) = log\,\left(\frac{N}{count\,(d \in D : t \in d)} \right)$$

Source: https://monkeylearn.com/blog/what-is-tf-idf/

In other words, the word *t* in the document *d* from the document set *D* is calculated as above.

TF-IDF evaluates how “relevant” a word is to document WITHIN a collection of documents.

Experimental setup. Describe how you are going to use the data to train and test the model. Explain how you will implement the model in the chosen framework and how you will judge the performance.

As previously stated, the GFC batch will be used as the training set. This means, the algorithms will be TRAINED upon this batch and then tested upon the batch of the post COVID-19 set to determine the sentiment of such terms and articles. The GFC batch was tagged with “positive” or “neutral” prior to inputting into the model.

Judging the performance of the model will be done by the accuracy score presented at the end of each model. This accuracy score will be the difference between the correct labels of each news article to the label the model tagged the article with. The greater the accuracy, the greater the model performed when tagging the news article with the correct sentiment.

What kind of hyper-parameters dd you search on? (e.g., learning rate)? How will you detect/prevent overfitting and extrapolation?

Of course, machine learning algorithms desire large sets of data in order to predict and acquire the knowledge of every different case in order to predict as accurately as possible. The distribution of positive and negative news articles were unbalanced, to say the least. As 2006 turned into 2007, the number of negative articles started to out-number the positive number of tagged-articles. The unbalanced data would likely have determined news articles to be skewed towards negative tagging. Thus, using SMOTE oversampling added positive examples to the training set so that the set had approximately the same number of negative and positive news articles. This method only applies to the training set as data leakage would result in a high accuracy scores that are unreliable.

Transforming each news articles into a specific vector (or set of features) was done by the tf-idf vectorizer. The features in the tf-idf approached were 5500 (most common words), thus each news article had 5500 features. Each feature was the tf-idf weight as described as above.

Once these modifications were complete, the training set was then fed into the algorithms.

Results. Describe the results of your experiments, using figures and tables wherever possible. Include all results (including all figures and tables) in the main body of the report, not in appendices. Provide an explanation of each figure and table that you include. Your discussions in this section will be the most important part of the report.

From the keyword extraction, four words appeared the most throughout all news articles:

1. Economy
2. Credit-Market
3. Mortgage-lending
4. Housing-market

5. Bear Sterns*
    a. This word did not actually appear as a top word, this was chosen from my inspection of the data

From these words, news articles that consisted of these news articles averaged to be tagged with either negative or positive the most.

	Keywords	Sentiment		Keywords	Sentiment
	Economy	Positive		Economy	Negative
	Credit Market	Positive		Credit Market	Negative
2006	Mortgage-Lending	Positive	2008	Mortgage-Lending	Negative
	Housing Market	Positive		Housing Market	Negative
	Bear Sterns	Positive		Bear Sterns	Negative
	Keywords	Sentiment		Keywords	Sentiment
	Economy	Positive		Economy	Negative
	Credit Market	Positive		Credit Market	Negative
2007	Mortgage-Lending	Positive	2009	Mortgage-Lending	Negative
	Housing Market	Positive		Housing Market	Negative
	Bear Sterns	Positive		Bear Sterns	Negative

	Keywords	Sentiment
	Economy	Negative
	Credit Market	Negative
2010	Mortgage-Lending	Negative
	Housing Market	Negative
	Bear Sterns	Negative

The Naïve Bayes Algorithm performed the best. The results of this model were followed through to the end:

	Keywords	Sentiment		Keywords	Sentiment
	Economy	Positive		Economy	Negative
	Credit Market	Positive		Credit Market	Positive
2019	Mortgage-Lending	Positive	2021	Mortgage-Lending	Positive
	Housing Market	Positive		Housing Market	Positive
	Keywords	Sentiment		Keywords	Sentiment
	Economy	Positive		Economy	Negative
	Credit Market	Negative		Credit Market	Positive
2020	Mortgage-Lending	Positive	2022 (present)	Mortgage-Lending	Negative
	Housing Market	Positive		Housing Market	Negative

Words such as *Economy*, *Mortgage Lending*, and *Housing Market* resulted in **negative** tagging by the end of each event. In contrast, *Credit Market* resulted in **positive** tagging by the end of the post-covid 19 event. These results may appear for a multitude of reasons. From my research, policies pertaining to the economic recovery by the government differed greatly from the GFC and the post COVID-19 recession. In 2009, government bailouts and TARP lending provided a cushion for banks to fall back on instead of collapsing. However, in 2021 – 2022, the government provided relief to individuals, families, and small businesses the most in order to bring consumption up through the economy and prevent one market character from being “shutdown” for some time.

Summary and conclusions. Summarize the results you obtained, explain what you have learned, and suggest improvements that could be made in the future.

In summary, sentiment analysis was used in this project to determine if a news article was positive or negative. The models were trained on data from the Global Financial Crisis (2007-2009) and tested on news articles surrounding the post-COVID 19 recession we are presently in.

These news articles contain words that are consistent in both batches of event and eventually are tagged with similar sentiment values before and after each event concludes.

Thus, continuing work on this project may be relevant for economists, data scientists, and anyone interested to understand how reporting and specific words talked about change as the economic fluctuates.

References. In addition to references used for background information or for the written portion, you should provide the links to the websites or github repos you borrowed code from.

https://colah.github.io/posts/2015-08-Understanding-LSTMs/

https://monkeylearn.com/sentiment-analysis/

https://towardsdatascience.com/a-guide-to-encoding-text-in-python-ef783e50f09e

https://www.aiperspectives.com/twitter-sentiment-analysis/

https://medium.com/@vasista/sentiment-analysis-using-svm-338d418e3ff1#:~:text=Sentiment%20Analysis%20is%20the%20NLP%20technique%20that%20performs%20on%20the,positive%2C%20negative%2C%20or%20neutral.

https://www.analyticsvidhya.com/blog/2021/06/natural-language-processing-sentiment-analysis-using-lstm/

https://www.analyticsvidhya.com/blog/2021/07/performing-sentiment-analysis-with-naive-bayes-classifier/

https://monkeylearn.com/blog/what-is-tf-idf/

Measuring Social Unrest Using Media Reports (Barret, Appendino, Nguyen, de Leon Miranda, 2021)

The Macroeconomic Impact of Social Unrest (Hadzi-Vaskov, Pienknagura, Antonio Ricci, 2020)

Measuring Economic Policy Uncertainty (Baker, Bloom, Davis, 2016)