Sentiment Prediction and Topic Modeling on Financial News Data

Jie Xu, Zijin Yu

2021 Fall

Project Narrative

Background & Stakeholders

Financial markets and financial news are closely related to each other (Alanyali et al., 2013). Alanyali et al. (2013) indicated that traders and corporates would proactively search for information reflecting their intended transactions and receive news from financial broadcasts or platforms which can also impose an influence on their financial behaviors. While financial news has a significant impact on the interpretation of the financial markets, it is important to take advantage of the great volume of financial news to simulate potential reactions by the receivers, thus characterizing major events that impact the market and recognizing future movements in the market. Prior studies have attempted to use machine learning in this field to portray financial movements. For example, Mahajan et al. (2008) proposed a text-mining system that analyzes financial news concerning the Indian stock market via the Latent Dirichlet Allocation (LDA) model by identifying topics and common words. For sentiment analysis, Schumaker et al. (2012) analyzed the choice of words and tone used by financial news, studied their correlation to the stock market, and made price predictions according to sentiments in financial news.

However, the problem we recognized is to properly model the behaviors of traders (Schumaker et al., 2012) in comprehensive methods as we found little literature combining sentiment analysis, topic annotation, and topic modeling regarding financial news. Therefore, this project works in the domain of finance and aims to predict positive, negative, and neutral sentiments in financial news headlines using machine learning, to categorize topics by manual annotation and prediction, and to perform topic modeling by unsupervised machine learning. Our intended audience for this project is corporations and firms that need to make profits by distinguishing behavioral patterns and predicting stock market changes. As stakeholders care about profits and benefits most, this project application of machine learning for solving the problem of lacking thorough financial news analysis can raise their interests by concise prediction.

Dataset Description

We try to build a news sentiment analysis model that can help to overcome the challenges of identifying the sentiments of the financial news. The data we worked on is the FinancialPhraseBank dataset, named “Sentiment Analysis for Financial News” from Kaggle. It is originally from a study, Good debt or bad debt: Detecting semantic orientations in economic texts, by Malo et al. (2014). It consists of two columns and around 5,000 news headlines. The columns present in the dataset are listed as follows:

> *sentiment: the polarity of the news headlines (positive, neutral or negative)*
> *text: the specific content of the news headlines*

Headline samples are paired with sentiment labels. We take the sentiment as the label for prediction, while the text corpus as the independent variables or features. Since this dataset only contains two columns, no column is set aside from modeling. Table 1 shows the first five rows of the dataset.

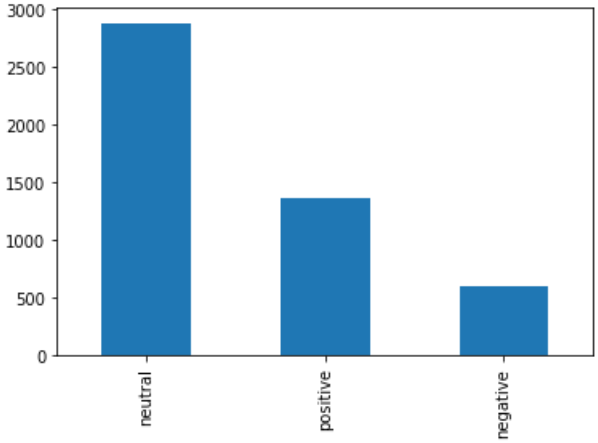
Table 1. Five Records of the Financial News Dataset

sentiment	text

neutral	According to Gran, the company has no plans to move all production to Russia, although that is where the company is growing.
neutral	Technopolis plans to develop in stages an area of no less than 100,000 square meters in order to host companies working in computer technologies and telecommunications, the statement said.
negative	The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility; contrary to earlier layoffs the company contracted the ranks of its office workers, the daily Postimees reported.
positive	With the new production plant the company would increase its capacity to meet the expected increase in demand and would improve the use of raw materials and therefore increase the production profitability.
positive	According to the company's updated strategy for the years 2009-2012, Basware targets a long-term net sales growth in the range of 20 % -40 % with an operating profit margin of 10 % -20 % of net sales.

For the content, news headlines differ in length and style. Texts could consist of words or combine words with numbers and percentages. Some samples contain irregular punctuation, double spaces or extra spaces around punctuation marks. Figure 1 shows the distribution of sentiments towards the news headlines. Neutral sentiment take the leading role in the whole dataset, while positive sentiments only appear as half of the amount, and negative appears even less, only more than 500 records.

Figure 1. Sentiment Distribution (Positive, Negative, Neutral)



Primary Task

Introduction

Sentiment analysis refers to identifying as well as classifying the sentiments that are expressed in the text source. News headlines can be good examples in generating a vast amount of sentiment data upon analysis. The opinion of the mass media can be easily read through in these headlines, which can also take lead in the public's opinion. Therefore, we are interested in developing an Automated Machine Learning Sentiment Analysis Model in order to compute the sentimental guide in news. In this project, we aim to analyze the sentiment of the financial news dataset by developing a machine learning pipeline comparing four classifiers: logistic regression, SGD,

SVC, and multinomial Naïve Bayes with Term Frequency- Inverse Document Frequency (TF-IDF). Metrics applied for performance measurement are confusion matrix and F1 Scores.

Methodology

Taking the original dataset, the dataset is split into a training dataset and a testing dataset at 20%. We set the pipeline for the modelling, from splitting the text using Count Vectorizer, transforming Dataset using TF-IDF Vectorizer, to fit the variables with a certain classifier. Then we train the model with the hyperparameter in the pipeline and evaluate the model with the test data. According to the result, we can tune the model into a better news sentiment prediction model, and then figure out the systematical error in the modeling process further analysis.

Experimental Setup

For a sentiment prediction modeling, we have to be cautious about overfitting, the case where model performance on the training dataset is improved at the cost of worse performance on data not seen during training, such as a holdout test dataset or new data. We can identify if the model has overfitted by first evaluating the model on the training dataset and then evaluating the same model on a holdout test dataset. Therefore we will use the *train_test_split()* function and split the data into 80 percent for training a model and 20 percent for evaluating it, which is about 4,000 samples for training and 1,000 for evaluating the model in this project.

Metrics

We try to build a model to identify the sentiments of the financial news. The following metrics are used to measure the modeling: accuracy score, checking the overall accuracy of the model with the division of number of Correct Queries and the total number of Queries; confusion matrix, telling how the model is performing with a combination of statistics including precision, recall, and f1-score. We are expected to see the model reaching an accuracy of 0.8, even better with a 0.9.

Features

As the original input was text, we have one more step to create machine learning-friendly input. One common approach is the *Bag of Words*, which functions by counting how many each word appears in certain documents (disregard of grammar and word order). We use the *CountVectorizer* method in the *sklearn* library. As the number of vocabulary is very large, we limit the size of the feature vectors to 1,000, and set the *ngram_range* and *max_df*. *ngram_range* means we cut one sentence by the number of ngram, and *ngram_range = (1,3)* means from unigrams to trigrams are included. Setting *max_df=0.8* helps to ignore terms that appear in more than 80% of the documents. We adjust the parameters to improve the result of accuracy.

Classifier

We basically run the model with the classifiers of Logistic Regression, Support Vector Machine and Naïve Bayes. With default settings, the classifiers perform differently based on the accuracy score, and Logistic Regression achieved a relatively high accuracy of 0.7. We then looked into the SGDClassifier, a generalized linear classifier using Stochastic Gradient Descent as a solver. With the classifier, we can use lots of different loss functions to tune the model and find the best SGD based linear model for training data. GridSearchCV() then help us to find the best solver to find the best hyper-parameters of *class_weight=balanced, loss=log, penalty=l2*.

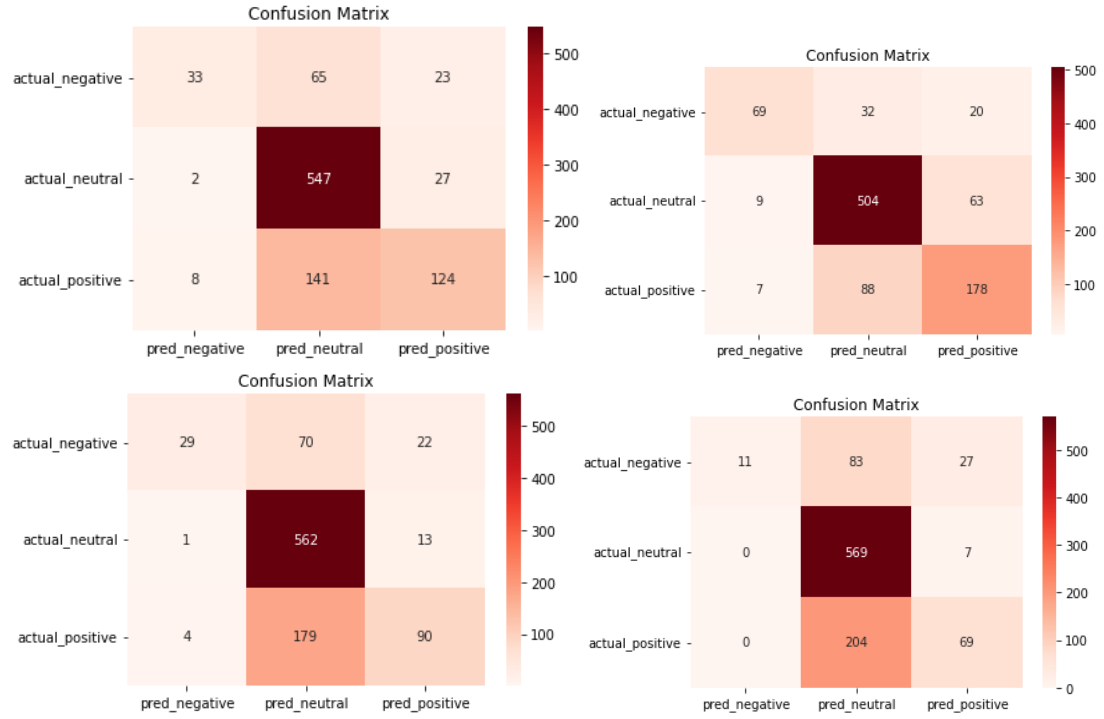
Results

Table 2 shows the main scoring result for new sentiment analysis with four different classifiers, including *LogisticRegresssion()*, *SGDClassifier()*, *SVC()* and *MultinomialNB()*. Detailed results for prediction of each sentiment class can be viewed in Figure 2.

Table 2. Results for News Sentiment Analysis with Different Classifiers

	Logistic Regression	SGD	SVM	Naïve Bayes
f1-score	0.40	0.67	0.37	0.17
accuracy	0.73	0.77	0.70	0.67

Figure 2. Confusion Matrix for News Sentiment Analysis with Different Classifiers

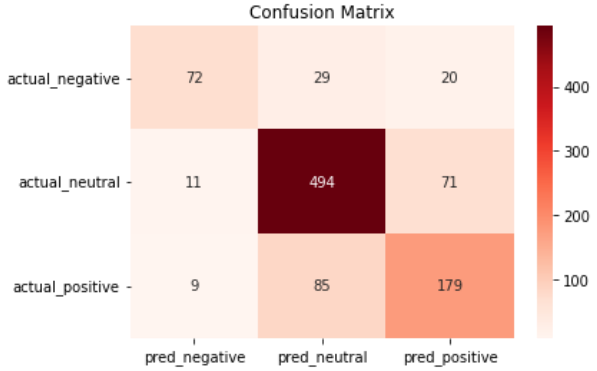


SGD Classifier performs the best among the classifiers. On this basis, we tuned the basic model with hyperparameters of *class_weight*, *loss* and *penalty*. The scoring of the best modelling is presented in the following table and figure, and the model achieves an accuracy of 0.77 with slightly higher f1-score and recall values on positive and negative sentiments.

Table 3. Tuning Results for News Sentiment Analysis with SGD Classifier

	precision	recall	f1-score	support
negative	0.78	0.60	0.68	121
neutral	0.81	0.86	0.83	576
positive	0.66	0.66	0.66	273
accuracy			0.77	970
macro avg	0.75	0.70	0.72	970
weighted avg	0.77	0.77	0.77	970

Figure 3. Confusion Matrix for Optimal Modeling



Error Analysis

The model performs well especially on prediction of neutral sentiments, while the prediction of positive, especially negative sentiments needs to be improved. The phenomenon led us back to the imbalance of the dataset. When the imbalanced classes were trimmed, we try to train the model on the balanced dataset and the accuracy was obviously lifted to 0.82. However, the model could not perform well on the original imbalanced dataset, and may need further detection if the real-world sentiments towards financial news are imbalanced.

Extension Task 1: Annotation

Task Definition

This extension task aims to design the new labeling scheme, *news topics*, through the annotation task. Based on common knowledge, I compiled an annotation manual, trying to impose definitions on the company's *finance news*, *product news*, and *structure news*. Identifying the insufficiency through sample dataset annotation, I revised the manual into a universal manual of company financial news annotation.

Motivation

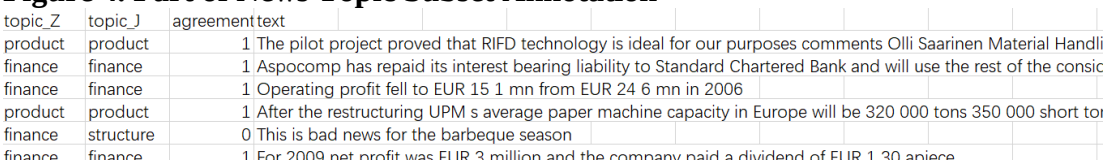
People in today's world are overwhelmed with data, which often leads to the problem of choosing something from a large set of options. Accordingly, various companies are targeting consumers with recommender systems, such as e-commerce giant Amazon and many traditional and innovative news media. Content topic prediction modelling can help identify the large amount of news contents into certain categories in a more efficient way. Based on the topic classification, companies can push certain content towards specific customer groups, which can greatly increase their marketing efficiency and add loyalty to their audience. This modeling can be representative of such mechanism, while the technology is of great potential in the real world.

Methods

The initial manual is to impose basic definitions on company's *finance news*, *product news*, and *structure news*, explained with specific examples. With the manual, I randomly picked 30 samples from the original financial news dataset, and asked my teammate to help annotate the subset. The percent agreement between raters of the sample dataset is 86.7%. Identifying the difference and broader discussion on topic classification with my teammate, I revised the

manual. With the new manual, I labeled another 400 instances into the three topics categories, and build a news topic prediction model.

Figure 4. Part of News Topic Subset Annotation



topic_Z	topic_J	agreement	text
product	product	1	The pilot project proved that RIFD technology is ideal for our purposes comments Olli Saarinen Material Handli
finance	finance	1	Aspocomp has repaid its interest bearing liability to Standard Chartered Bank and will use the rest of the consic
finance	finance	1	Operating profit fell to EUR 15 1 mn from EUR 24 6 mn in 2006
product	product	1	After the restructuring UPM s average paper machine capacity in Europe will be 320 000 tons 350 000 short to
finance	structure	0	This is bad news for the barbeque season
finance	finance	1	For 2009 net profit was EUR 3 million and the company paid a dividend of EUR 1 30 apiece

Figure 5. Part of News Topic Annotation

sentiment	topic	text
neutral	finance	Atria will also buy the shares of Kauhajoen Teurastamokiinteistot Oy Kauhajoki slaughterhouse property from Itikka Co oper
neutral	structure	The Group brand portfolio includes the leading brand in the industry Rapala and other global brands like VMC Storm E
positive	finance	Ragutis controlled by the Finnish brewery Olvi achieved a 5 7 percent rise in beer sales to 22 6 million liters and held a 10
positive	product	It provides customers with industry leading elevators escalators and innovative solutions for maintenance and modernizatic
positive	structure	efficiency improvement measures 20 January 2010 Finnish stationery and gift retailer Tiimari HEL TII1V said today that it w
neutral	structure	Apart from Nordea also Ergo is competing for the position among the top three pension funds providers in Estonia

Results

The accuracy of the model on testing data turned out to be 0.6. Additionally, I input three sample data for the model and two of them turns out to be correctly classified. The accuracy of the model still remains to be improved.

Extension Task 2: Unsupervised Learning

Task Definition

This extension task aims to perform topic modeling as a type of unsupervised learning to extract essential topics from the financial news dataset using the Latent Dirichlet Allocation (LDA) model. I structured my experiment so that I had a held-out test set, which was not part of any of the experiments and tuning of the cluster analysis. By setting the test size as 0.2, I split the dataset into 20% testing data and 80% training data. Three main steps included are python implementation, data preprocessing and cleaning such as punctuations, stopwords removal, and lemmatization, and LDA modeling. During the data preprocessing, I checked the frequencies of the top weighted words in the dataset and visualized them in a bar plot.

Motivation

This additional task is interesting as it can general detailed insight into a large body of financial news in an automated way. It allows specific tagging of financial news by indicating various levels of relevance of topics, and thus gives the readers a hint on ongoing and future transactions. Stakeholders such as firms will be interested in this task since it can provide a detailed breakdown of topics or realms from a large amount of financial news. Clustering can be useful in grouping similar news articles automatically and acting as a basis for predictive modeling (Topic Modeling, n.d.). This unsupervised learning can affect the way the audience perceives the primary and the first extension task as it works as a comparison between manual topic annotation and topic modeling and automatic labeling. The differences of specific divisions of topics between both manual work and machine learning will provide new insights into financial news analysis.

Methods

Data Preprocessing: Punctuations & Stopwords Removal, Lemmatization

Figures 6, 7 and 8 show corresponding bar plots in each preprocessing step. In the first visualization, the most frequent words are stopwords such as “the”, “and”, “of”, and “in”, as well as punctuations like “,” and “.” Since these words are less important, I removed all punctuations, numbers, stopwords and words composed of less than two letters. The second bar chart shows improvements after removal: the most common words are shown as “EUR”, “the”, “company”, “Finnish”, “sales”, “million”, “profit”, etc, which are more relevant to the field of financing. However, some words appear to be variations of the same root or derived from the same word such as “Finnish” and “Finland”. Lemmatization is thus introduced to further reduce noise from the text by extracting the proper lemma of each word and therefore reducing multiple forms to a single word lemma. I tokenized the news headlines and then lemmatized the tokenized texts. In addition, I kept only nouns and adjectives to reflect the most important part of each sentence. After revision, the most common words included “company”, “sale”, “year”, “profit”, “net”, “Finnish”, “service”, and “market”.

Figure 6. Original word frequencies

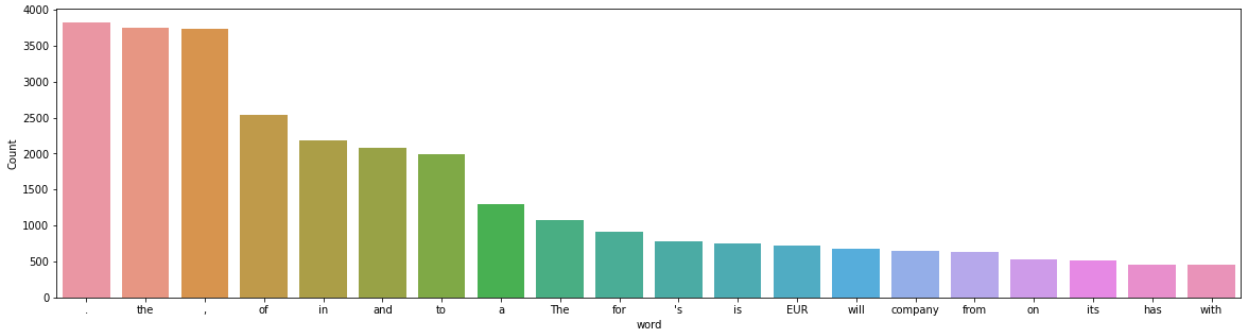


Figure 7. Word frequencies after stopwords removal

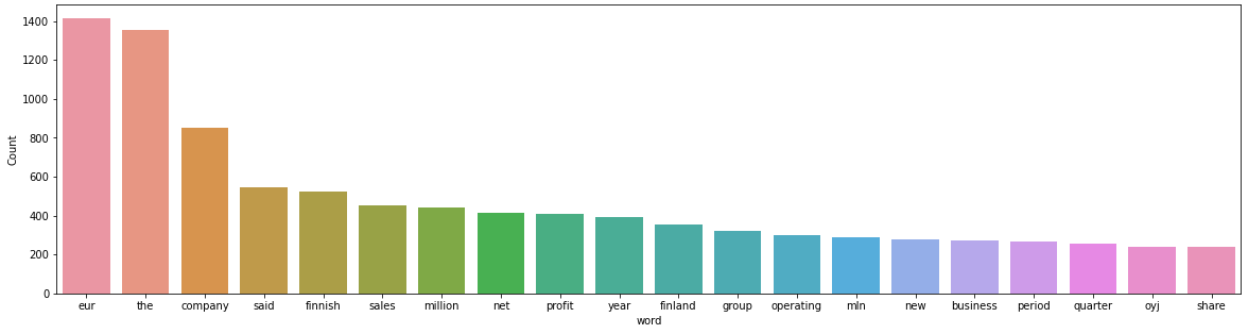
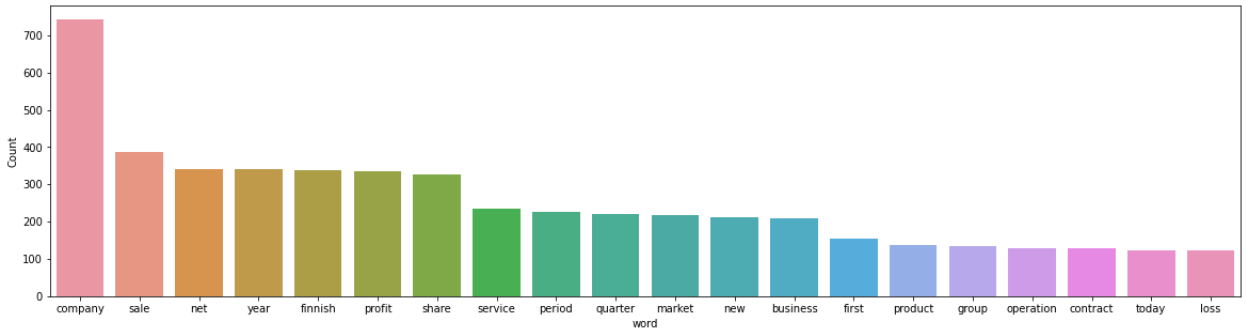


Figure 8. Word frequencies after lemmatization



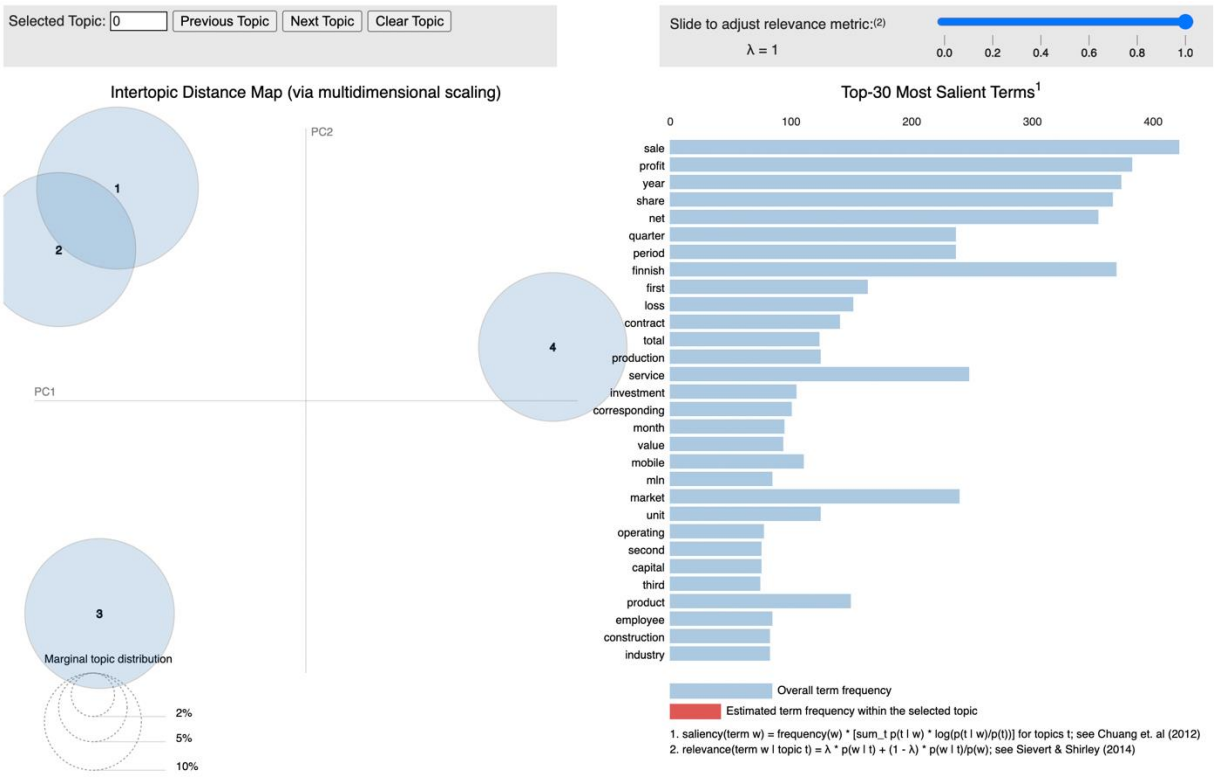
Clustering Algorithm: LDA Modeling

An LDA model was built according to the result of data preprocessing. Based on the results and outputs from the LDA model, I word labeled these topics by their relevant departments. Based on the keywords, I divided them into four categories: *structure*, *profit*, *product*, and *market*. For Topic 1, I labeled it as a *structure* since it contains high frequency words such as business, company, operation, agreement, today. Topic 2 was labeled as *profit* as it contains keywords like net, year, sale, profit, year, net, quarter, loss, month, min, revenue. As for Topic 3, it contains keywords like company, business, customer, service, solution, production, hence I labeled it as *product*. Topic 4 was being labeled as *market* as the higher weightage words include share, company, total, market, investment, value, capital, and stock. The distribution of topics and their corresponding word weights can be seen in the following:

(0,
'0.033"company" + 0.023*"finnish" + 0.014*"service" + 0.013*"percent" + 0.013*"business" + 0.013*"agreement" + 0.010*"operation" + 0.009*"sale" + 0.009*"today" + 0.008*"financial"'),*
(1,
'0.046"net" + 0.046*"year" + 0.045*"profit" + 0.044*"sale" + 0.030*"quarter" + 0.030*"period" + 0.028*"company" + 0.024*"finnish" + 0.021*"first" + 0.017*"loss"'),*
(2,
'0.020"company" + 0.017*"business" + 0.016*"product" + 0.015*"customer" + 0.013*"service" + 0.012*"new" + 0.011*"value" + 0.011*"order" + 0.010*"mobile" + 0.010*"production"'),*
(3,
'0.060"share" + 0.027*"company" + 0.018*"market" + 0.017*"new" + 0.012*"capital" + 0.012*"number" + 0.009*"total" + 0.009*"option" + 0.009*"right" + 0.008*"earning"')]*

The weights after each word are demonstrated in a decreasing trend, reflecting the level of significance a keyword has on the single topic. For example, the top ten keywords that constitute the first topic are “company”, “Finnish”, “service”, “percent”, “business”, “agreement”, “operation”, “sale”, “today”, and “financial”. The weight of “company” on this topic is 0.033. Figure 9 is an interactive map in Python visualizing the distances among topics in a two-dimensional space and demonstrating each topic with its top frequent words.

Figure 9. LDA topic modeling results



Results

Model Measurement, Hyperparameter Tuning, & Supervised Labelling

To measure the performance of the model, I used perplexity and coherence score as metrics. The perplexity score was -7.076 and the coherence score was 0.349. For hyperparameter tuning, I need to see the topics to know if the model makes sense or not. The approach I adopted to find the most proper number of topics is to build many LDA models with different numbers of topics from three to six to understand which amount makes sense. To make a practical application, I made automatic labeling by determining which topic is best for each news headline. The topic with the highest percentage distribution for each headline was identified as the main topic. For example, in Figure 5, the first two rows were identified as of the first topic, structure. This extension task helps the audience better understand the primary task by indicating which sentiment tends to be the most common or most related to each topic.

Table 4. Automatic Labeling According to LDA Model

	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	1.0	0.4587	net, year, profit, sale, quarter, period, company, finnish, first, loss	According Gran the company has plans move all production Russia although that where the company growing
1	1	1.0	0.8725	net, year, profit, sale, quarter, period, company, finnish, first, loss	Technopolis plans develop stages area less than square meters order host companies working computer technologies and telecommunications the statement said

Reference

Dataset: Sentiment Analysis for Financial News, Kaggle, https://www.kaggle.com/ankurzing/sentiment-analysis-for-financial-news

Alanyali, M., Moat, H., & Preis, T. (2013). Quantifying the Relationship Between Financial News and the Stock Market. *Sci Rep 3*, 3578. https://doi.org/10.1038/srep03578

Mahajan, A., Dey L., & Haque, S. M. (2008). Mining Financial News for Major Events and Their Impacts on the Market. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 423-426. https://doi.org/10.1109/WIIAT.2008.309

Malo, P., Sinha, A., Takala, P., Korhonen, P., Wallenius, J. (2014). Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. *Journal of the American Society for Information Science and Technology, 65.*

Schumaker, R. P., Zhang, Y., Huang, C., & Chen, H. (2012). Evaluating sentiment in financial news articles. *Decision Support Systems, 53*(3), 458-464. https://doi.org/10.1016/j.dss.2012.03.001

Topic Modeling: Summarizing Financial News (n.d.). Applied AI. Retrieved December 17, 2021 from https://ml4trading.io/chapter/14