# CS305 Final Project: Predicting Market Sentiment Using Financial News Headlines

Fall 2023

Crystal Luo, Victoria Lu, and Bridget Sheng

**Background, Motivations, and Research Questions**

The goal of our project is to predict the financial sentiment of news headlines. While judging a single headline's sentiment may be simple, it is much harder to assess the overall market sentiment as there are hundreds and thousands of news articles released every day. There are other sentiment indicators out there, but they often do not directly measure market sentiment. Thus, we try to use machine learning to bulk-predict the sentiment of financial news headlines. This will allow us to understand the overall market sentiment more quickly and better, and this has a few important implications:

1) Market sentiment is a driver of asset prices. Some optimism creates more investor demand, but excessive optimism may cause an asset to be overvalued.
2) Market sentiment often reflects broader economic conditions.
3) Market sentiment is also important for policymakers. If the market responds poorly to a new policy (e.g., stimulus), this may be a sign that the policy is insufficient.

Given this context, our research questions are:

1) How accurately can we classify the sentiment of financial news headlines?
2) What factors are the most important in the process?

We care about both predicting correctly and interpreting. The former is crucial for the practical use of the model, while the latter helps build intuition. As it is not a trivial task to constantly scrape news from the web and run machine learning models on them, we believe that having an intuitive understanding of what words/phrases contribute the most to predicting sentiment is useful.

**Data Description and Related Literature**

Our dataset Sentiment Analysis for Financial News on Kaggle contains the sentiments for 4837 financial news headlines from the perspective of a retail investor. The variable of interest is *Sentiment*, which is a categorical variable. It is annotated by 16 finance researchers/students and has a subjective label of "positive," "negative," or "neutral." The other column has a variable-length news headline (e.g., "The international electronic industry company Elcoteq has laid off tens of employees from its Tallinn facility."). These headlines make up a random subset of 10,000 articles from the LexisNexis database, which have a good coverage across small and large companies in different industries listed on the Nasdaq Helsinki (a stock exchange located in Helsinki, Finland), as well as different news sources. Figure 1 is a Word Cloud of all the headlines.

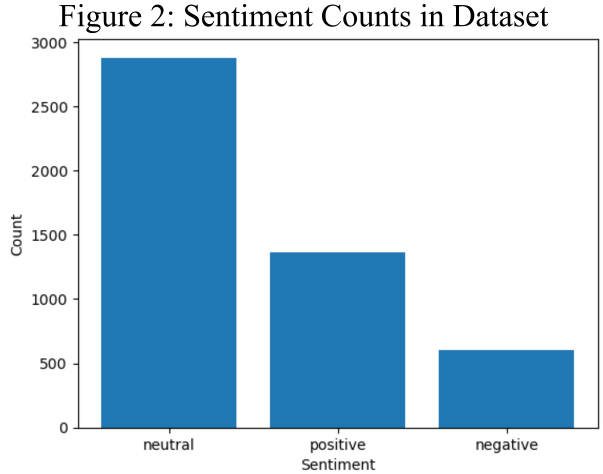
The dataset has been used by a few research projects. The best-performing model we saw used the language model BERT and was able to achieve a weighted-average precision, recall, and F1-score of 85% across the board. Some academic papers further explored non-ML models such as LPS and W-Loughran. In the paper "Good debt or bad debt: Detecting Semantic Orientations in Economic Texts," the authors were able to achieve a 70%-80% accuracy, recall, precision, and F1-score on all classes using the LPS approach.

Figure 1: Word Cloud for News Headlines


**Class Imbalance**

One potential source of bias in our dataset is that there is a class imbalance (neutral: 59%, positive: 28%, negative: 12%; see Figure 2). Given this, it is likely that our ML models will do better at correctly identifying the neutral class compared to the other two classes. To assess the performance of our models more fairly, instead of accuracy, we focused on the precision [=True positives/ (True positives + False Positives)] and recall [=True positives / (True positives + False Negatives)] of individual classes. In our context, a higher precision for a class means that when predicting on that class, more predictions are correct, while a high recall for a class means that the model can capture more of the actual instances for that class.

Figure 2: Sentiment Counts in Dataset




Since we are dealing with three classes, we also computed an average precision and recall weighted by the number of observations in each class. Furthermore, we do not think that there is a difference in the cost of making a Type I vs. Type II error in our context (i.e., incorrectly predicting a positive sentiment to be a negative or neutral one is no worse than incorrectly predicting a negative sentiment to be a neutral or positive one), so we calculated the weighted

F1-score [=2*(Weighted precision * Weighted recall) / (Weighted precision + Weighted recall)], which accounts for the natural tradeoff between precision and recall.

**Featurization and Data Splitting**

We used a Bag of Words approach to convert the textual dataset into a numerical one, which works by first breaking up all the headlines into words, then counting the number of times each word occurs, and finally normalizing the counts so that less frequently occurring words are given larger weights. Our featurized unigram dataset has the dimension 4837 x 8972, where the cell corresponding to row i, column j counts the number of times the word j appears in headline i. Our featurized bigram dataset has the dimension 4837 x 49,802, which considers, in addition to individual words, all adjacent two-words combinations. For both datasets, we shuffled the data and then randomly split the data into 20% testing and 80% training.

**Models and Hyperparameter Tuning**

Given the large feature space of our dataset, it is not clear whether the datapoints have linear or non-linear decision boundaries. Therefore, we tried all models that are suitable for a multiclass-classification problem, including logistic regression, support vector machines, decision tree, random forest, kNN, perceptron, and neural networks. For each model, we first tuned hyperparameters that we deemed important using 5-fold cross-validation optimized on weighted F1-score via a coarse-to-fine grid search approach, then fitted the model with the optimized parameters on the training set, and finally evaluated model performance on the testing set. Table 1 displays the hyperparameters we tuned for each model, the best value selected, and our rationale for tuning them.

Table 1: Hyperparameters for Models (Unigram and Bigram)

| Model | Hyperparameter | Best value (unigram) | Best value (bigram) | Rationale & Description |
|---|---|---|---|---|
| Logistic regression | C (Regularization parameter) | 20 | 90 | To avoid overfitting/find the right balance between minimizing cost and adding more terms to the model; a smaller C puts a heavier penalty on added terms. |
| SVM with a linear kernel | C (Regularization parameter) | 5 | 10 | To avoid overfitting/find the right balance between allowing for a smooth decision boundary and minimizing misclassification rate; a larger C penalizes misclassifications more heavily. |
| SVM with a polynomial kernel | C (Regularization parameter) | 20 | 10 | Same as for SVM with a linear kernel |
| | Degree | 2 | 2 | To adjust the shape of the decision boundaries; higher degrees are associated with more complex decision boundaries. |
| SVM with a RBF kernel | C (Regularization parameter) | 20 | 30 | Same as for SVM with a linear kernel |
| | Gamma | 0.1 | 0.1 | To adjust the variance of the underlying Gaussian function; a larger gamma is associated with a Gaussian function with a smaller variance, which leads to more complex decision boundaries. |

| | | | | |
|---|---|---|---|---|
| kNN | k (Number of neighbors to consider) | 13 | 9 | To find the right balance between bias and variance. A smaller k results in a more flexible model but may be more sensitive to noise. |
| Decision tree | max_depth | 30 | 30 | To prevent the tree from being too deep and thus overfitting. |
| Random Forest | max_depth | 500 | | Same as for decision tree |
| Perceptron | NA | NA | NA | NA |
| Neural networks | hidden_layer_sizes | (100, ) *One layer of 100 neurons | (100, ) | To avoid overfitting; more hidden layers can capture more abstract relationships but may be less generalizable to new data. |
| | activation | logistic | logistic | To adjust the shape of the decision boundaries. |
| | alpha (Regularization parameter) | 0.01 | 0.01 | To prevent overfitting; A higher alpha value places heavier penalty on larger weights in the network. |

*We were unable to tune the hyperparameters for neural networks with the bigram approach because the algorithm took a very long time to run (running the model one time takes around 3h, so tuning 36 combinations of the 3 hyperparameters using 5-fold CV would take about 540h). We were also unable to perform PCA due to our X_train matrix being sparse. Therefore, we decided to use the same hyperparameters as those optimized for the unigram approach.

### Model Performance

Tables 2 and 3 summarize the performance of our models for the unigram and bigram approach. Overall, for the unigram approach, logistic regression and neural networks, which also uses the logistic activation function, showed the best performance across the board. For the bigram approach, the logistic regression model had the best performance. One surprising observation is that models with linear decision boundaries (e.g., logistic regression, perceptron, and SVM with a linear kernel) appeared to generally outperform ones with non-linear decision boundaries (e.g., SVM with polynomial and rbf kernels, decision tree, and random forest). This may indicate that even though there are many features, the underlying relationships between these features and sentiment are more-or-less linear.

Table 2: Model Performance (Unigram)

| **Model** | **Weighted average precision** | **Weighted average recall** | **Weighted average F1-score** |
|---|---|---|---|
| Logistic regression | 0.76 | 0.76 | 0.76 |
| SVM with a linear kernel | 0.75 | 0.75 | 0.75 |
| SVM with a polynomial kernel | 0.75 | 0.74 | 0.71 |
| SVM with an rbf kernel | 0.75 | 0.74 | 0.71 |
| Decision tree | 0.66 | 0.68 | 0.66 |
| Random forest | 0.76 | 0.74 | 0.71 |
| kNN | 0.67 | 0.69 | 0.67 |
| Perceptron | 0.74 | 0.75 | 0.75 |
| Neural networks | 0.76 | 0.76 | 0.76 |

Table 3: Model Performance (Bigram)

| Model | Weighted average precision | Weighted average recall | Weighted average F1-score |
|---|---|---|---|
| Logistic regression | 0.77 | 0.78 | 0.77 |
| SVM with a linear kernel | 0.76 | 0.76 | 0.76 |
| SVM with a polynomial kernel | 0.75 | 0.71 | 0.66 |
| SVM with an rbf kernel | 0.75 | 0.71 | 0.66 |
| Decision tree | 0.71 | 0.71 | 0.70 |
| Random forest | 0.77 | 0.75 | 0.72 |
| kNN | 0.67 | 0.68 | 0.67 |
| Perceptron | 0.76 | 0.76 | 0.76 |
| Neural networks | 0.77 | 0.77 | 0.77 |

**Model Interpretation**

Out of all the ML models we employed, logistic regression, decision tree, random forest, and perceptron are interpretable. For logistic regression, we were able to find the words/phrases that correspond to the most positive and negative coefficients (see Appendix 1). Overall, for the unigram approach, words that are the most indicative of sentiment are directional verbs like "increased," "rose", and "up" (for positive sentiment), and "decreased," "down," and "fell" (for negative sentiment). This is mostly consistent with intuition, but a caveat is that while a word like "decrease" can be linked to earnings decreasing, which is negative for companies, it can also be associated with expenses decreasing, which is positive for companies. The fact that we see an overwhelmingly positive association between the direction indicated by the verbs and sentiment indicates that reporters or retail investors may place a higher emphasis on revenue generation as opposed to cost cutting. Another pattern that is worth noting is that many words with large positive coefficients (indicative of negative sentiment) are related to employment, with some examples being "laid," "staff," "cut," and "jobs." We thought that this is interesting because while negative news about a company can come in various forms, it seems like news about layoffs are viewed particularly negatively. Moreover, even though layoffs may hurt a company's public image and lead to higher expenses in the short-term, it should not be totally negative to an investor, as it means lower labor costs for a company in the medium-to-long term. As a result, we thought that the pattern of job words being strongly associated with negative sentiment may indicate that retail investors have some short-term bias. Moving from the unigram to the bigram approach, we did not see any meaningful changes in which words/phrases were weighted most heavily. For the most part, only prepositions were added to the original words (e.g., "rose" became "rose to").

For random forest, we were able to generate an importance plot based on the Gini importance measure (see Figures 3 & 4). The results largely agree with those of the logistic regression. One slight difference is that words without much meaning (e.g., "to", "from", and "in") played a larger role in the random forest model compared to logistic regression, which may explain why the random forest model's performance was not as strong. There was also minimal change in words/phrases identified as important moving from the unigram to bigram approach.

Figure 3: Importance Plot for RF (Unigram)

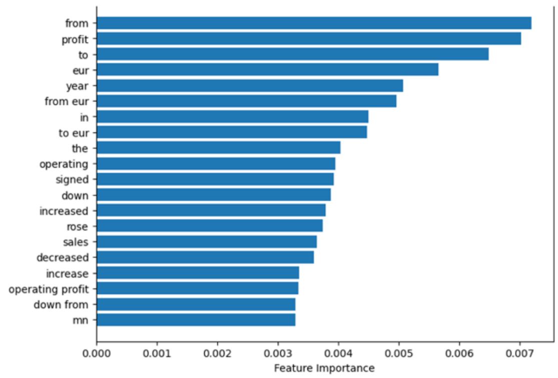
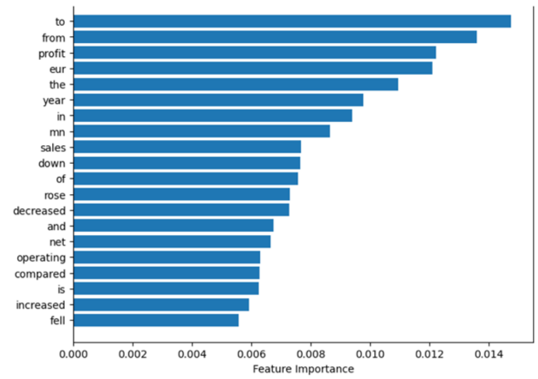



Figure 4: Importance Plot for RF (Bigram)




As for perceptron, it has very similar highly weighted words compared to logistic regression. As before, the bigram approach also did not add much intuitive insight.

With all three models above, something anomalous that we noticed is that certain numbers like 20, 28, and 51 were weighted heavily (e.g., the phrase "28 points" had the fourth most-negative weights in perceptron). We reexamined our dataset and determined that this is likely due to chance as we were unable to find any special patterns in the headlines involving these numbers in the dataset.

**Best Model, Limitations, and Extensions**

Comparing the bigram approach to unigram approach, we saw minimal improvement in the models' performances. Given that the bigram approach adds almost 40,000 more columns to the dataset, which slows down computational speed significantly, we think that the unigram approach is sufficient. Further considering all models, we prefer the logistic regression model the most, which has a weighted precision, recall, and F1-socre of 76%, because it has the best performance across the board, is very interpretable, and has a fast computational speed. We think that the performance of this model is fairly strong considering its complexity relative to other models utilized by many researchers. Nevertheless, we think that there are several limitations to our approach and best model.

One concern is that there is some subjectivity involved in the annotation of the sentiment label. The fact that the dataset was annotated collectively by 16 researchers with a background in finance helps reduce subjectivity, but the concern is still present. We tried to use K-means clustering to assess how subjective the annotations were. We divided the headlines into 3 clusters, based on their distances apart. With the unigram approach, when compared to the actual sentiments, the match was only about 14%. However, after we switched to the bigram approach, the match rose to 62%. Since k-means clustering is a form of unsupervised learning, we cannot definitively tell whether the clusters are created based on sentiment (i.e., it can be based on the number of words, company names, topic discussed, etc.), but the elbow plots for both the unigram and bigram approaches indicated that dividing the deadlines into 2-4 clusters is optimal (see Appendix 3), which corresponds well with the sentiment label. Assuming that the clusters were formed based on sentiment, the fact that the match rate was 62% when considering just two-words combinations suggests that while there may have been some subjectivity, it was likely not a huge concern.

Another limitation of our approach is external validity. The dataset is about companies in Finland, which make up only a small part of the world financial market. Because of this, our models may not perform well when applied to companies in other countries or financial news that are not company related. Given the class imbalance, our models are also likely to do better when the actual sentiment is neutral compared to positive or negative.

As an experiment, we scraped 289 additional news headlines from Reuters from the period 2019-04-01 to 2019-04-07. Based on the Word Cloud (see Figure 5), these headlines cover a much wider range of finance topics from company earnings to the economy and trade but are mostly related to the US. We featurized this dataset using the unigram approach, established the common features between this dataset and our original dataset, filled in the features exclusive to our original dataset with 0s, predicted the sentiments, and finally assessed subjectively whether these sentiments actually match the news headlines (see Appendix 4 for sample predictions). Overall, while we agreed with most of the positive and negative predictions, we disagreed with many of the neutral predictions. An overwhelming percentage of the news headlines were classified as neutral as opposed to positive or negative (15 positives, 48 negatives, and 227 neutrals). Upon further examination, we determined that many of these neutral headlines should probably be classified as positive or negative. This could be due to the fact that many features were filled in as zeroes.

Figure 5: Word Cloud for Reuters News Headlines


We think that it is quite hard to address the issue of external validity when using supervised machine learning models. The solution would be to expand to a more comprehensive headlines dataset, but to fit supervised machine learning models, it is necessary to annotate by hand the sentiment labels, which can be very laborious. Alternatively, one may try to assemble the dataset by matching each headline to a stock or economic indicator by keyword and then use the change in these data as a proxy for sentiment. Additionally, to address the issue of class imbalance, it may be worth it to see if balancing the training set using balancing techniques like undersampling, oversampling, or synthetic minority oversampling improves model performance.

Works Cited

Brown, G., & Cliff, M. T. (2001). Investor sentiment and asset valuation. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.292139

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2013). Good debt or bad debt: Detecting Semantic Orientations in Economic Texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796.

Memocan. (2023, December 13). *Sentiment unveiled: BERT & Elmo explored* 🚀. Kaggle: Your Machine Learning and Data Science Community. https://www.kaggle.com/code/memocan/sentiment-unveiled-bert-elmo-explored

*Sentiment analysis for financial news*. (n.d.). https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news

Appendices

## Appendix 1: Logistic Regression Weights

Unigram:

| Top 20 words indicative of positive sentiment | Top 20 words indicative of negative sentiment |
|---|---|
| up: -5.298449723740245 | down: 9.700829444406512 |
| rose: -4.828971269188565 | decreased: 9.469335052485558 |
| increased: -4.367824761457641 | fell: 8.223989977694727 |
| increase: -3.8268634721255665 | declined: 5.440689718615271 |
| new: -3.4282549658444497 | off: 5.056024567565487 |
| flight: -3.0010794822057902 | result: 4.958881279456801 |
| 20: -2.7797343546459827 | below: 4.85686538152675 |
| business: -2.595444261765905 | dropped: 4.729798165505634 |
| will: -2.489714158121455 | lower: 4.593780748773688 |
| started: -2.441806449123444 | lay: 4.502054970545271 |
| all: -2.373675409439433 | slipped: 4.445295545777649 |
| an: -2.2925880932091696 | staff: 4.358753805242649 |
| program: -2.2399543488735736 | reduction: 4.038549742746596 |
| and: -2.2266809081499472 | cut: 4.00694873030839 |
| technology: -2.170020322762406 | warning: 3.977473017975769 |
| euros: -2.1547903701598408 | gone: 3.8333770065339725 |
| improved: -2.1519477754695324 | layoffs: 3.7439441048907702 |
| approximately: -2.1446525933399956 | jobs: 3.7220540737599466 |
| transferred: -2.122191664772566 | because: 3.6665707341755733 |
| annual: -2.1182433417198654 | burdened: 3.5285927355277407 |

Bigram:

| Top 20 words/phrases indicative of positive sentiment | Top 20 words/phrases indicative of negative sentiment |
|---|---|
| was loss: -4.525012797114654 | down: 9.949186553060787 |
| rose: -4.320706250177454 | decreased: 9.483507531320237 |
| up: -4.199602888568563 | fell: 8.49499832969549 |
| increased: -4.166697887639045 | off: 5.810208129386943 |
| rose to: -3.7354283577714122 | lower: 5.542165850651252 |
| up from: -3.5700867933926586 | result: 5.5188276313321785 |
| increase: -3.501075077674327 | decreased to: 5.267851284981941 |
| and: -3.284067356885426 | staff: 4.802719921584133 |
| new: -3.1550033748303425 | down from: 4.786685951662783 |
| is: -2.6970096754288133 | declined: 4.690910505145163 |
| business: -2.6937142526330815 | lay: 4.6738097522324 |
| profit rose: -2.6761040960963074 | dropped: 4.227789171807733 |
| will: -2.643197016643059 | to profit: 4.080460637245008 |
| an: -2.6381016312683885 | cut: 3.754322653272699 |
| to negative: -2.588993974442831 | fell to: 3.6819810043240118 |
| eur0 01: -2.51328050148955 | was negative: 3.641511394739218 |
| period was: -2.494247344210332 | reduction: 3.6329506426690186 |
| grew: -2.400821217844131 | below: 3.5624567854318405 |
| mn up: -2.1935224604988774 | profit warning: 3.5431131041957533 |
| increased by: -2.19304720703623 | warning: 3.540579244380214 |

Appendix 2: Perceptron Weights

Unigram:

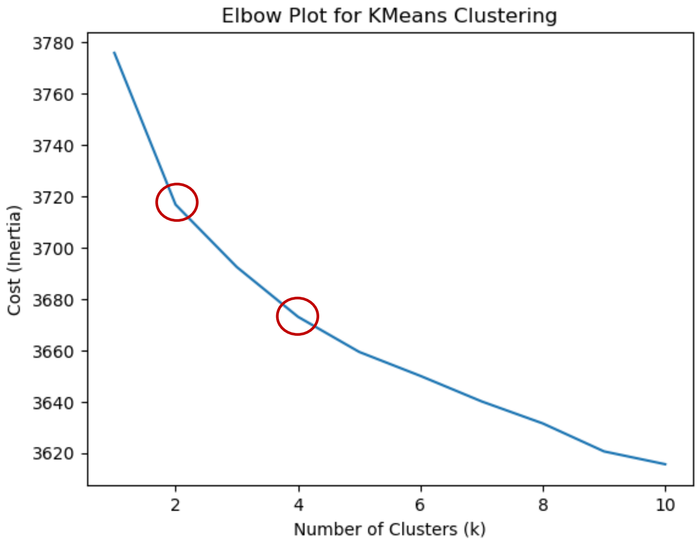
| Top 20 words indicative of positive sentiment | Top 20 words indicative of negative sentiment |
|---|---|
| rose: -2.856506185103066 | down: 4.253202297397022 |
| increased: -2.466191604160188 | decreased: 3.639872923422945 |
| up: -2.269764187077253 | fell: 3.2800641823232155 |
| 20: -1.9172804703958841 | declined: 1.9656809803928732 |
| increase: -1.7295132596137277 | drop: 1.907524652790269 |
| improved: -1.6199330777782055 | result: 1.9012814760796894 |
| flight: -1.6139559298475832 | dropped: 1.8831920395589368 |
| euros: -1.5207836036372486 | slipped: 1.8642085207823127 |
| operations: -1.409913407830428 | burdened: 1.835443078395734 |
| 442: -1.3175402465899744 | warning: 1.8079894600854631 |
| business: -1.3064067615507575 | longer: 1.7186551956316123 |
| 2011: -1.2734526482897057 | gone: 1.6990433395378148 |
| 159: -1.235754603373918 | below: 1.6417670664791357 |
| plc: -1.2310427770239596 | off: 1.6273440522635756 |
| new: -1.2180002503433887 | given: 1.604804437584269 |
| includes: -1.2069983881090343 | because: 1.5995608903748513 |
| program: -1.1803486211863208 | strike: 1.4984977855329287 |
| applicant: -1.1735384890287066 | parent: 1.463716895397079 |
| started: -1.1571982656033697 | decline: 1.4600361506853612 |
| it: -1.1522266310724514 | kroons: 1.4352617961088339 |

Bigram:

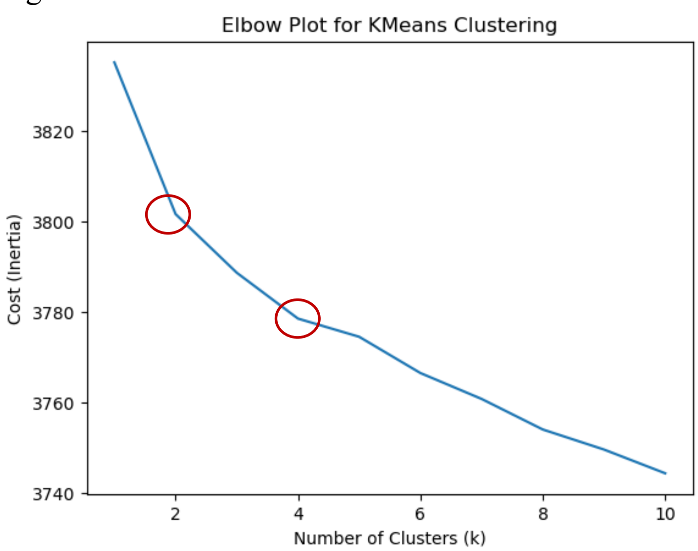
| Top 20 words indicative of positive sentiment | Top 20 words indicative of negative sentiment |
|---|---|
| rose: -1.294477709400527 | down: 2.2344543312979077 |
| rose to: -1.21681662129432 | decreased: 2.1245584819715306 |
| 28 points: -1.0920576189343176 | fell: 1.790737057473698 |
| rose 28: -1.0920576189343176 | mn in: 1.3298333901437867 |
| increased: -1.0782046992981065 | lower: 1.3220589031379586 |
| up from: -1.0555676743073175 | decreased to: 1.2571484334492873 |
| was loss: -1.0377121900828046 | down from: 1.2465780109720805 |
| up: -1.0084523913101004 | off: 1.13976057002856 |
| period was: -0.9376054167130201 | to profit: 1.0869564075155482 |
| increase: -0.8236608894000695 | was negative: 1.0372903860571603 |
| profit rose: -0.8096862741048911 | eur0 05: 1.0300697911790073 |
| to negative: -0.7733953413840993 | result: 1.0234421391997603 |
| program: -0.7618812322771833 | cut: 1.0035429829532154 |
| expects: -0.7525369885168669 | staff: 0.959048103709246 |
| nordea: -0.7231014214432812 | below: 0.9371401421038408 |
| to loss: -0.6827456174491195 | because of: 0.9344935121061244 |
| improved: -0.6806055580143467 | lay: 0.9284739242461951 |
| the loss: -0.6637603440335667 | profit fell: 0.8803529341381022 |
| percent: -0.6546000569133472 | 36 points: 0.8727780029637465 |
| it: -0.6541557297911257 | rose 36: 0.8727780029637465 |

Appendix 3: Clustering Elbow Plots

Unigram:




Bigram:



Appendix 4: Reuters News Headlines Sample Predictions

Neutral:

```
neutral----Ousted Nissan boss Ghosn's video to be shown Tuesday: Kyodo
neutral----KPMG plans overhaul of British business: The Times
neutral----New NAFTA deal 'in trouble', bruised by elections, tariff rows
neutral----American Airlines extends 737 MAX cancellations through June 5
neutral----Fiat Chrysler to pay Tesla hundreds of millions of euros to pool fleet
```

Positive:

```
positive----Carlyle agrees to buy 30 percent stake in Spain's Cepsa: FT
positive----Copper producers gather; electric cars seen driving demand growth
positive----Hyundai Motor denies tie-up with Tencent on driverless car software
positive----Oil prices rise 1.5 percent as strong U.S. economic data eases demand concerns
positive----Trump tries fresh approach with long-delayed Keystone XL pipeline
```

Negative:

```
negative----Big banks to report first quarter results with lowered expectations
negative----Boeing to reduce 737 production in wake of MAX crashes: statement
negative----U.S. jobless claims hit 49-year low; labor market resilient
negative----U.N. agency works to clamp down on illicit shipping practices
negative----Deutsche Bank bans staff from Dorchester hotels after Brunei implements ho
mosexuality laws
```