➢ Dataset and the Goal of the Study

- Dataset 3: 400,000 reviews on Amazon and their corresponding positive or negative.
- Goal: Construct learning algorithms to classify reviews as positive or negative
- Such analysis in Machine Learning (ML) is called Sentiment Analysis.
- Examples of Positive and Negative Reviews:

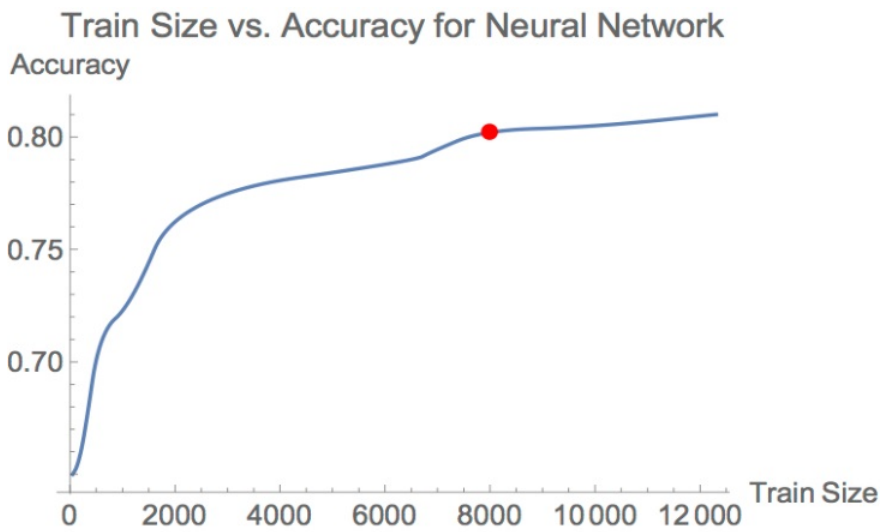
positive|Great CD: My lovely Pat has one of the GREAT voices of her generation. I have listened …
negative|DVD Player crapped out after one year: I also began having the incorrect disc problems …

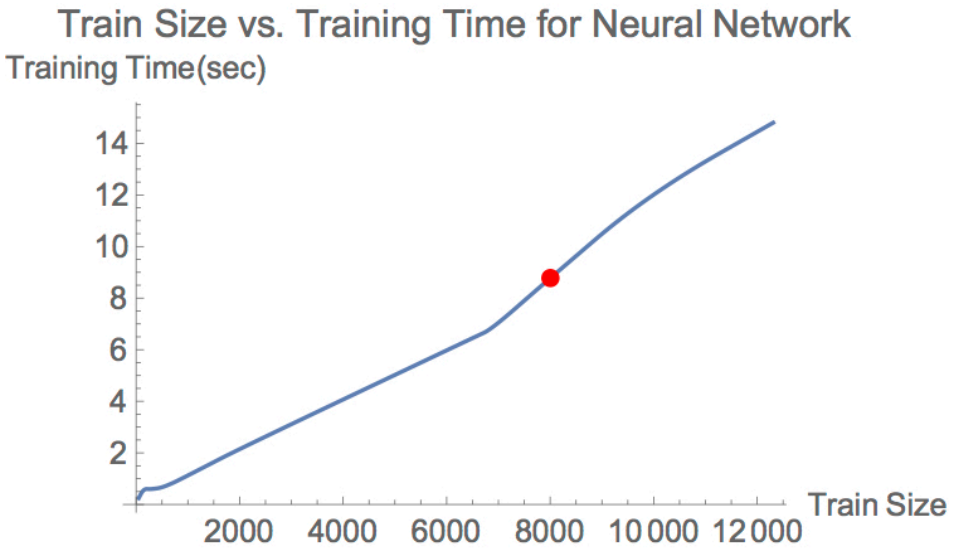
➢ Train and Test Sets

- Training Set and Test Set contain 80% and 20% of the dataset, respectively.
- Dataset is divided into two sets such that every 5th sample belongs to test set and the remaining samples belong to training set.

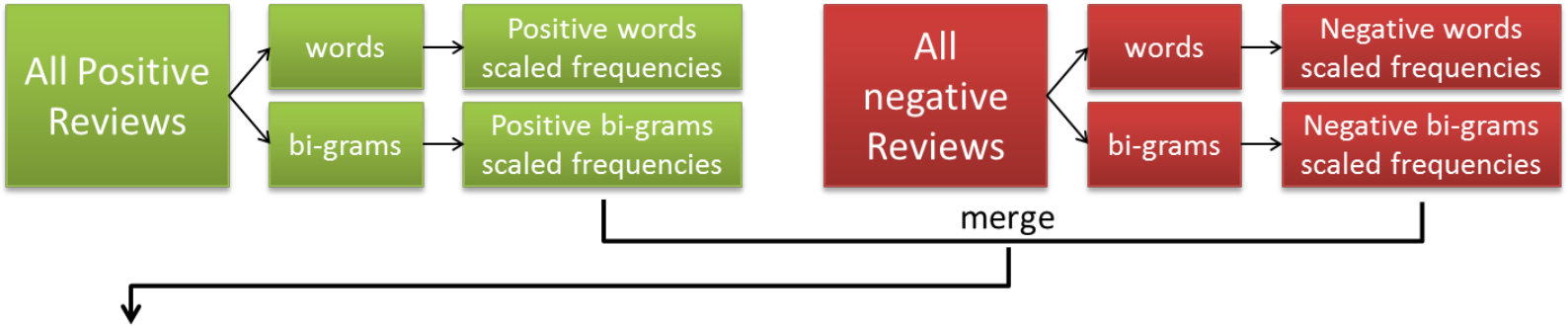
➢ Selection of the Train and Test Sets

- Trade-off between computational time and performance.
- Selecting a subset of the train and test sets for our analysis.
- Train subset contains 8000 samples distributed uniformly over the original train set.
- This number is chosen using the following analysis.





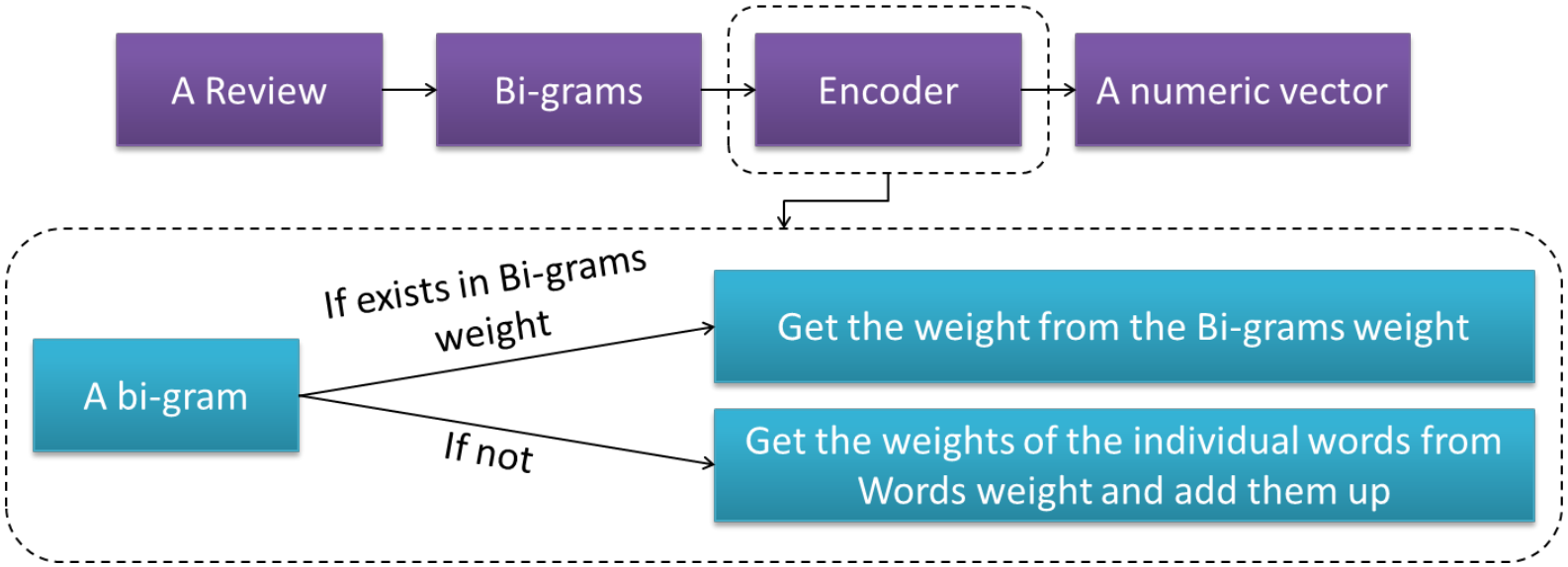
➢ Main Approach for Text Preprocessing



Words weight = {great -> 53.55, love -> 22.45, ….. , bad -> -13.59, waste -> -14.21}

Bi-grams weight = {(highly, recommend) -> 4.01, ….., (waster, money) -> -5.87}

➢ Encoding Reviews



➢ Classifications

▪ Main Approach

- Defined the encoder function converts texts in reviews into vectors of numbers.
- Used the function as a Feature Extractor in the **Classify** function in Mathematica.
- Performed classifications using 3 algorithms, below:

1. **Artificial Neural Network (ANN):** Mandatory

2. **Random Forest (RF):** Mandatory, ensemble variant of Decision Tree

3. **Support Vector Machines (SVM):**

   - Widely used in the real-world text classification problems.
   - Able to classify problems that may not be linearly separable using non-linear kernel (e.g. RBF).
   - Less susceptible to noise and outlier points as presented in practical ML experiments.
   - However, one downside of using SVMs is large running time when dealing with large datasets.

➢ Classifications

▪ Alternative I

- Used original reviews in text and their corresponding labels to construct a classifier using **Markov** algorithm.
- Classify function using Markov Method can handle texts in the reviews automatically with no preprocessing.
- Not efficient for other algorithms, such as Neural Network, Random Forest, etc.

▪ Alternative II

- Each review is converted to a list words (lower case and without stop words).
- Used FeatureExtraction function with "TFIDF“ and "DimensionReducedVector" options to extract features.
- Performed classification by Classify function with **Logistic Regression (LR).**
- Not Efficient for other algorithms, such as Neural Network, Random Forest, etc.

➢ Results and Visualizations

▪ Performance Table : Accuracy, Precision, Recall and Area under ROC Curves

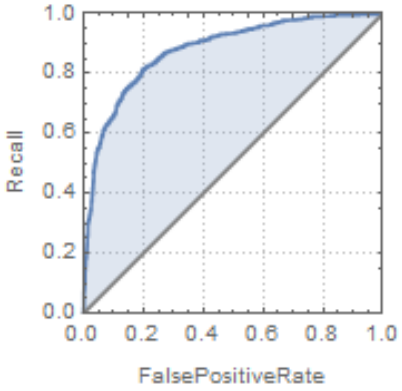
Methods		Accuracy	Precision(pos)	Precision(neg)	Recall(pos)	Recall(neg)	ROC Area(pos)	ROC Area(neg)
Main Approach	ANN	0.800	0.813	0.788	0.770	0.827	0.879	0.879
	RF	0.798	0.816	0.782	0.760	0.834	0.858	0.850
	SVM	0.801	0.817	0.787	0.767	0.833	0.880	0.880
Alternative I	Markov	0.837	0.859	0.817	0.798	0.874	0.901	0.902
Alternative II	LR	0.837	0.841	0.832	0.823	0.850	0.907	0.907

- All 3 methods in Main Approach result in classifiers with similar performance (e.g. accuracy around %80).
- Random Forest produced lower values for area under ROC curve metric, which indicates lower performance comparing to ANN and SVM.
- Alternative I (Markov) and Alternative II (Logistic Regression) result in classifiers with higher performance (e.g. accuracy) and yet relatively efficient.

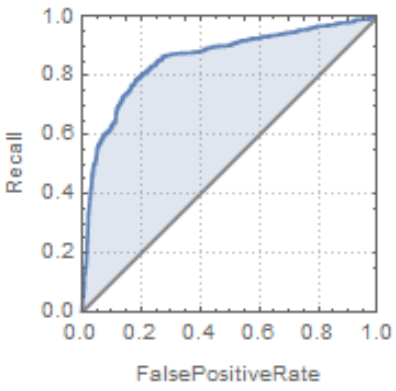
➢ Results and Visualizations

▪ ROC Curves

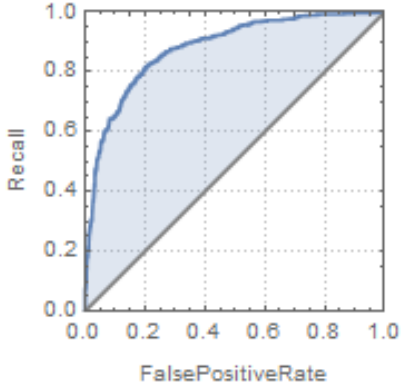
Main Approach: ANN



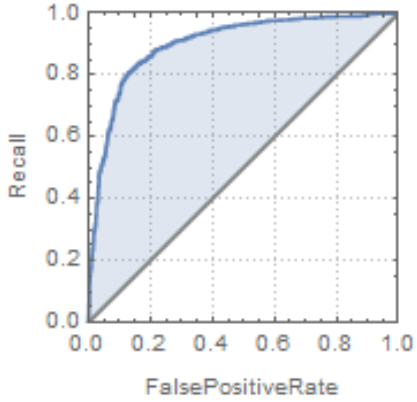
Main Approach: RF



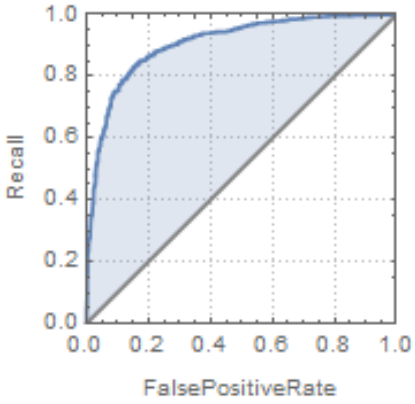
Main Approach: SVM



Alternative I: Markov



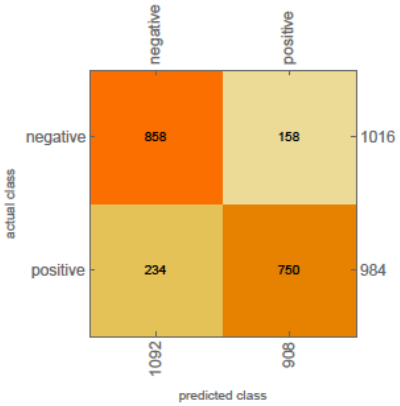
Alternative II: Logistic Regression



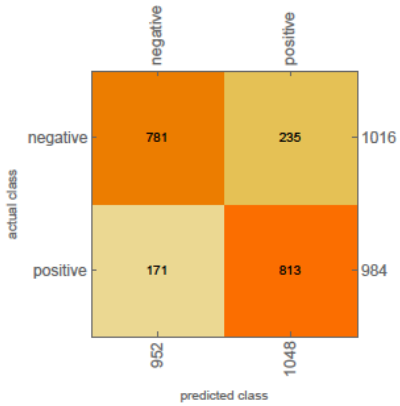
➢ Results and Visualizations

▪ Confusion Matrix Plot

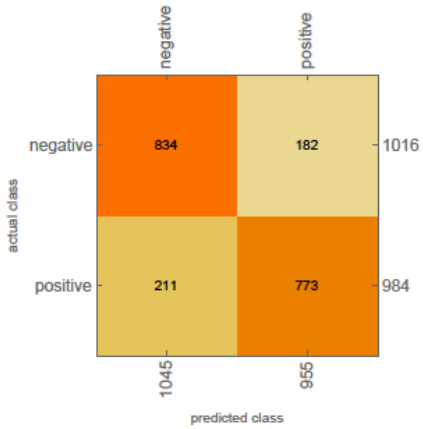
Main Approach: ANN



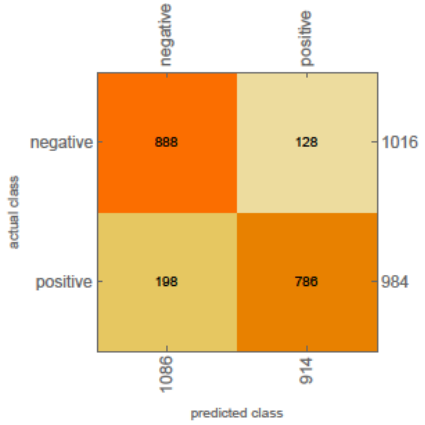
Main Approach: RF



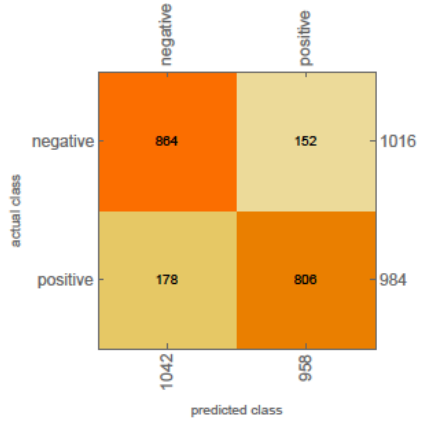
Main Approach: SVM



Alternative I: Markov



Alternative II: Logistic Regression



➢ Conclusions and Recommendations

- All 3 methods in Main Approach result in classifiers with similar performance (e.g. accuracy around %80).
- Among three methods used in the main approach, SVM and RF result in the highest and lowest values for area under ROC curve metric.
- Alternative I (Markov) and Alternative II (Logistic Regression) result in classifiers with higher performance (e.g. accuracy) and yet relatively efficient.
- Alternatives I and II are not efficient when using algorithms, such as Neural Network, Random Forest, etc.
- Note that the above conclusions are limited to the configuration considered in this study.

- Increasing the number of training set can enhance the accuracy of the classifier, especially when the features are selected such that they can predict the labels accurately.
- Increasing the number of features (But, if too large causes overfitting problem)
- Cross-Validation techniques such as k-fold cross-validation.
- Ensemble Machine Learning methods (combining the results of multiple learners).