Textual Analytics Scoring and Validation

Avneet Pal Kour, Chandrika Amarkhed, Merin Joy, Paras Sethi, Sri Ram Sagar Kappagantula

Volgenau School of Engineering, George Mason University, Fairfax, Virginia

ABSTRACT

Principal Financial Group (PFG), a global investment management leader, offers class retirement services, insurance solutions and asset management services to businesses, institutional investors and individuals preparing for the future. To achieve the company's mission, management needs to take various decisions as per financial statements, company records and market conditions. However, evaluating these large number of financial and non-financial documents is a painstaking task. This project aims to build a 'Textural Analytics Classifier' that classifies financial unclassified text to give meaningful insights that the management can use to make important decisions on growth, opportunity, strategy and stability.

The project utilizes the data derived from financial sources like analyst notes, stock narratives, transcribed announcements, emails, etc. synthesized by Decooda, a text analytics service partner of Principal Financial Group. The annotated data received from Decooda is preprocessed, followed by training of 4 different models on 8 different financial word lists, to create the baselines. The best contributing features from each trained model are then selected using three python machine learning feature selection methods. These top features are then combined and trained to give the overall best features and best results. The results are visualized using PowerBI dashboards.

METHODOLOGY

Planning and Analysis
Understanding Client needs, gathering information and document annotated information

Research Papers
Finding the datsets, wordlists, methods and algorithms used in reaserch papers

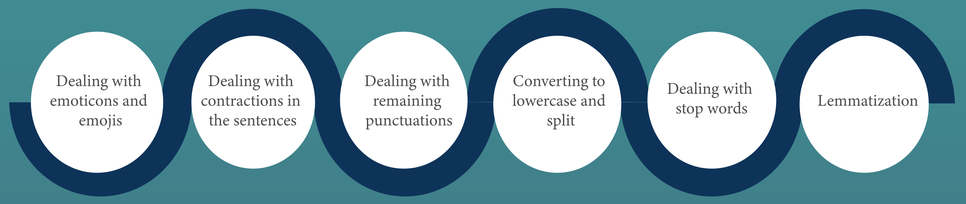
Implementation
Pre-processing, designing and hypothesis testing

Modeling
Models and techniques used in reaserch papers

Results
Predictive model results using PowerBI dashboards

DATASETS

DATA PRE-PROCESSING



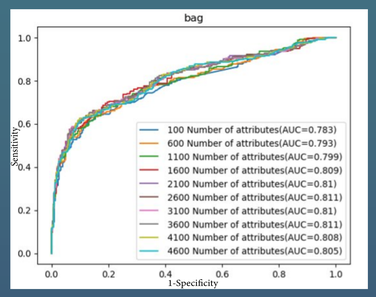
FEATURE SELECTION

Mutual_info_classif
Measures the mutual dependency between the variables. Estimates mutual information for a discrete target variable.

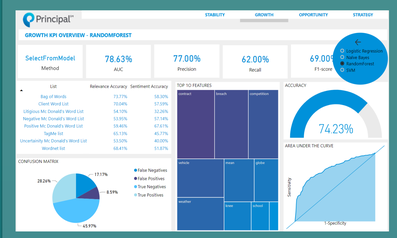
SelectFromModel
Meta-transformer for selecting features based on importance weights.

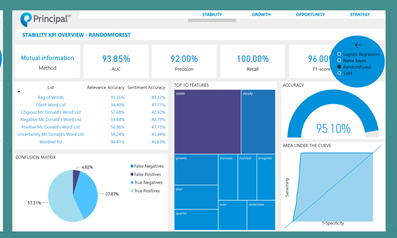
Chi-square
Measures dependency between stochastic measures. "Weeds out" the features that are the most likely to be independent of class and therefore irrelevant for classification. Compute chi-squared stats between each non-negative feature and class.

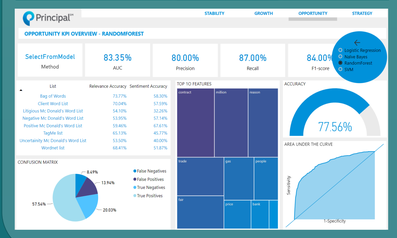
Example
Following is an example of chi-square feature selection method performed on the word list Bag of Words. Using Area Under the Curve (AUC) value we determine the number of top attributes and best features.

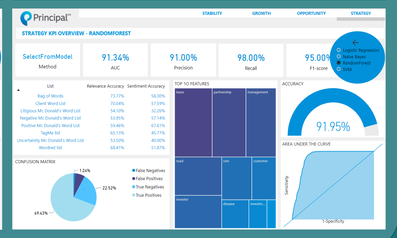


RESULTS









CONCLUSION

Including feature selection improves model results considerably compared to the baseline models. In this case, feature selection is also responsible for dimentionality reduction of more than thousands of features. Secondly, by using lists like Word net and Bag of Words we were able to increase the interoperability of our models. These models are not just limited to a specific dataset anymore. Also, results achieved using various natural language and feature selection techniques demonstrate a direct relationship between features and labels. In most of the cases, Random Forest model performs better compared to other modeling algorithms.

FUTURE WORK

Models can be trained and tested with much larger datasets as the data provided after annotations was not large and diverse enough. Word embeddings such as glove or word2vec can be used for feature engineering. Also, neural networks and advanced deep leaning models can be implemented.

ACKNOWLEDGEMENT

Special thanks to Professor F. Berlin, James Baldo, Rajesh Aggarwal and the Teaching Assistant Manan Grover for guiding and mentoring us throughout the project timeframe. Also, thanks to Principle Financial Group (PFG) and Talha Naushad for showing confidence in our skills and sharing valuable data and resources that helped in the completion of this project.

REFERENCES

[1] Steven L. Heston, and Nitish Ranjan Sinha, April 23, 2014," News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns."
[2] Samuel B. Bonsall IV & Brian P. Miller, 20 March 2017," The impact of narrative disclosure readability on bond ratings and the cost of debt."
*other References are provided in our Final Report.