Stock Portfolio Management: Prediction of Risk using Text Classification

Abhishek Sanwaliya, Kripa Shanker and Subhas C. Misra

***Abstract*—Decision making task utilizes different categorization techniques which have major contribution in building automated system capable to automate the decision for better organization and management of resources. The objective of this research is to assess the relative performance of some well-known classification methods for managing portfolio by predicting financial performance of a company on the basis of text documents. This approach is quite useful for managing portfolios on the basis of predicted risk category by framing the risk categorization task to assess company's financial outlook in the form of a classification problem. Availability of financial text information in electronic form and positive correlation between news reports on company's financial status make its classification a useful tool for investors to predict financial outlook of a company before investing their money in particular stock. Major contribution of classifier is that it enables to manage investor's portfolio. Among the proposed classification techniques combined naïve bayes and K-NN (NBKNN) with bigrams and unigrams with Latent Semantic Analysis (LSA) approach for features extraction has high potential of classifying and predicting risk from available financial news. We developed a text classification algorithm that classifies financial news article by using a combination of a reduced but highly informative word feature sets and a variant of weighted majority algorithm. Short learning time is additional advantage of this technique resulted due to its computational efficiency along with relatively high accuracy. We measure the accuracy on precision basis and also do a comparative analysis of combined NBKNN with other classification techniques. The major contribution lies in developing new classification method and its application in portfolio management.**

***Index Terms*—Naïve Bayes, K-NN, Bigrams, Unigrams. Portfolio Management.**

I. INTRODUCTION

Availability of large source of information about financial position and performance of a company makes it necessary for investors to integrate and manage this information for effective decision making. Stock portfolio management is the area which needs aid of automated decision support system (DSS). In order to manage stock portfolio we propose an intelligent system which can assist user to evaluate the risk associated with the individual company of interest on the basis of extracted terms and classification technique used. Our decision support system work in intelligent and effective manner which is helpful at the conclusive end for the users for different reasons like performance history, stock price, earnings per share and risk associated with individual holdings in their stock portfolio. This evaluation leads the system to proactively intimate the user about associated risk level with its tolerance limits towards risk. The conventional way is to access the news agent who gathered financial news from available online news providers Reuters, CNN financial network etc. Our automated DSS helps user to retrieve the important contents from available news articles, evaluate or classify them on the basis of classification technique and finally able to present the information in a meaningful and conclusive manner. So we have classifier as basic and most important component of our DSS. We develop a combined classification algorithm which is capable to classify news article into "High risk potential", "Medium risk potential", "Low risk potential" and "no risk". Our methodology focuses primarily on: (1) Extracted Features in a particular category which should be well distinguished and highly informative in nature which can allow a classifier to estimate the category of a news article with high probability; (2) classification algorithm based on prominent features selection process among identified features to perform well for this problem. Briefly, the proposed technique predicts the label of a news articles through voting process among identified feature set. The feature set is defined as a set of words describes a particular category relatively well and accordingly help a classification algorithm discern the boundary of a class (e.g. "high risk potential") from that of another. We made use of two types of feature sets: a set of words highly frequently co-occurred in a particular class and a number of word clusters semantically coherent. Our problem can be identified as classification problem in which essentially we have to assign appropriate class to each unlabelled article on the basis of semantic content of the document. Classification techniques used for discussion are: naive bayes [7],[3], KNN [6], centroid based approaches [1],[6],[14],[18], decision tree (DT) [13],[5],[15] and rocchio [10]. Current classification problem deals with relatively objective and confined classes like "High risk potential" and "no risk" for a company's financial outlook as compare to classification of news articles into "politics" or "economics".

The text classification has several characteristics that make it a difficult domain for the use of machine learning, including a very large number of input features, class noise, and a large percentage of features that are irrelevant. The paper is organized as follows. Section 2 will give the overview of our task in terms of the text classification context. Section 3 describes pre-processing of text. In Section 4, we describe procedure of feature set preparation for specific classification with consideration of the company's

Manuscript received November 30, 2009.

Abhishek Sanwaliya is with the Indian Institute of Technology, Kanpur, India (phone: 91-9452044881; fax: +91-512-2597376; e-mail: abhisan@ iitk.ac.in; abhisanwaliya@gmail.com).

Dr. Kripa Shanker is Vice Chancellor of Uttar Pradesh Technical University, Lucknow, India. He is also with the Department of Industrial and Management Engineering, Indian Institute of Technology, Kanpur, India as senior faculty member (e-mail: ks@iitk.ac.in).

Dr. Subhas Chandra Misra was Research Scientist at Harvard University, USA. He is with the Department of Industrial and Management Engineering, Indian Institute of Technology, Kanpur, India as faculty member(e-mail: subhasm@iitk.ac.in).

financial outlook. Section 5 elaborates training and testing algorithm for classification. Section 6 provides the experimental results and compares them with those of existing methods. Section 7 discusses the results and the future work respectively.

II. Categorization of Financial News Articles

Concisely, we develop an algorithm to classify each given news article into predefined classes on the basis of refereed company's financial status. The financial news articles gathered for experiments were manually labeled into 5 classes by considering how explicitly they mentioned the company's financial status. We categorize five categories to classify news articles:

Table 1. Table captions should be placed above the table

Class	Remarks	Examples
High Risk Potential (HRP)	News scripts which refer to predictions of future losses, or no profitability and News scripts which show bad evidences of the company's financial status explicitly.	Certain: Shares of company X fell in early Bombay Stock exchange Uncertain: Company X warned last month that First Quarter results could fall short of expectations.
Medium Risk Potential (MRP)	News scripts which refer to predictions of future profitability, and Forecasts.	Certain: Company X recovered from previous losses but still have to gain its original pace Uncertain: ABC and XYZ Inc. announced plans to develop an industry initiative.
Low Risk Potential (LRP)	News scripts which did not mention anything about the financial wellbeing of the company explicitly.	Certain: Marginal increase in profit along with hope as MoU signed yesterday Uncertain: ABC Company predicts fourth-quarter earnings will be high
Zero Risk Potential (ZRP)	News scripts which show good evidences of the company's financial status explicitly.	Certain: Shares of ABC Company rose by 2 percent from $12 to $14.4
Neutral (N)	News article that do not belongs to any category and difficult to determine the company's current financial status.	Owner of ABC company purchase his exclusive flat in posh area of London.

Each class is subdivided into certain and uncertain category. Certain gives information relevant to actual happening while uncertain provides prediction. We classify articles on the basis of definite or predicted financial news sentiments of particular company.

III. Text Preprocessing

In order to produce efficient results with high accuracy first we exclude the terms which are semantically insignificant. Table 2 includes pre-processing steps.

Table 2:Pre-Processing of Text

Normalization	To obtain a uniform text we adopt normalization in which we convert text to lowercase so that the distinction between uppercase and lowercase is ignored.
Tagging	Part-of-speech (POS) tagging is the process of assigning a part-of-speech like noun, verb, pronoun, preposition, adjective or other lexical class marker to each word in a sentence. It is based on "The Penn Treebank Tag set" [12]
Tokenization	Tokenization is the process of reducing a message to its colloquial components.
Dimensionality Reduction	Dimensionality reduction is a process to reduce the space of a document. Removal of the non-context words which occur with very high frequency in most documents and do not carry any semantic meaning for categorization and hence are insignificant in making distinction among different documents. A classifier should be tuned to the real characteristic of the training data. It increases accuracy and decreases the learning time.
Stemming and Lemmatization	Stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. The *Porter stemming algorithm*[1] (or '*Porter stemmer*')[9] is a process for removing the commoner morphological and in flexional endings from words in English. The wordnet lemmatizer only removes affixes if the resulting word is in its dictionary. This additional checking process makes the lemmatizer slower than the above stemmers. Stemming most commonly collapses derivationally related words, whereas lemmatization commonly only collapses the different inflectional forms of a lemma.

Combined effect of text filtration yields feature set having huge potential to produce efficient and effective classification. Dimension reduction provides improved efficiency by selecting relevant terms to prepare feature set. As a result of stemming and lemmatization effective feature

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

set prepared in order to achieve higher efficiency. Pre filtration has major impact on classification accuracy.

IV. Feature Set Preparation

A feature set is defined as a set of words (or phrases) specifies a particular class and accordingly help a classification algorithm discern the boundary of a class from that of another. We used bigrams, extracted from text, as contents of feature set. However we use this terminology because we focused on identifying a good set of word feature, rather than building a good classification algorithm. We made use of two types of feature sets: (1) Bi-grams approach: a set consisting highly frequently co-occurred with a particular class (2) Unigram Approach: a word cluster with similar semantic context. They are based on the concept of word co-occurrence.

A. Bi-gram Feature Selection

A co-occurred phrase is a word pair which that frequently occurred in a typical sequence in documents belongs to the same class. It is not necessary that they follow same sequence but should follow syntactic sequence of nearby words. It enables to prepare well distinguished co-occurred phrases which are strongly associated with its class and having potential of discriminating among available different categories. Another facet is to select terms having high frequency in order to reduce the noise of data. Complexity in classification increases with increase size of feature set. For efficient and prominent classification As well as t, however, is not easy to select such a set of word pairs due to the inherent complexities a strong association between a bigram and a class is necessary. A number of bigrams are initially compiled after removing stop-words. To determine a strong association between a bigram and a particular class, the information gain measure was employed [4].To determine the characteristic of financial news we concentrate on sentences containing companies name explicitly and then consolidate whole information to determine the class. Correlation between bigrams and class clearly determines the category of financial news.

B. Unigram Feature Selection

Another method for identifying feature set is unigrams approach which utilizes clustering of words into groups of similar concepts. The word similarity is estimated by co-occurrence between two words in sentence. Word similarity index provides good approximation of co-occurrence. In this approach we measure the similarity by computing cosine angle between two word vectors. This provides insight that if documents are having more number of co-occurred word higher would be the similarity value. We applied the concept of Latent Semantic Analysis (LSA) due to lack of semantically similar group detection ability of conventional weight based methods. To capture semantic coherence we use Latent Semantic Analysis [11]. We represent our feature set as original word document matrix to capture the semantic coherence of the text. Then we extract most important single factor to measure the covariance of inverted matrix. This way LSA captures the "semantic" value of given text document. It becomes easier to trace the similarity of documents using LS as it represents text in subjective way as compared to conventional approach in which all weight goes to high frequency terms. A word in the identified vocabulary is represented a word vector. A hierarchical agglomerative clustering [8], [2] is employed to group words. Unigram enables to prepare feature set which can provide better results.

V. Classification of News by NBKNN

Financial news data set obtained after text pre-processing has comparatively reduced dimensions. Processed text is the collection of terms remained after text filtration. After dimension reduction we still select only those terms which are discriminative in nature for efficient classification. This objective is achieved by either of two methods: (1) Bigrams feature selection (2) Unigrams feature selection method in order to make a feature–set. After selecting feature-set terms we provide relevant feature-set for both NB and KNN classifier. On one hand NB's feature–set contains terms and their respective frequency in order to calculate probability score whereas KNN's feature set is prepared by terms and their associated weight for calculating the Euclidian's length which is calculated using vector prepared from term weight obtained from discriminative term extraction technique. Afterwards training of both classifiers using their respective feature set is performed. Testing of unlabelled articles is done when classifier is supervised through training feature-set. Testing provides measure of score associated with unlabelled article in order to make conclusive classification. As objective of KNN is to assign a particular class to unlabelled article from which its distance is minimum when both labeled and unlabelled articles represented as vector. Similarly for NB we assign category to unlabelled article having maximum probability score. Hence we define the objective function which minimizes KNN score and simultaneously maximizes NB score which are the prime requirements of both classifiers and when combined they produce a score which assist categorization. To sum up we try to minimize the score of KNN and maximize the NB score and then combine both of them in order to categorize the article to that class for which it has minimum value of combined score obtained from combined Naïve Bayes and K-NN (NBKNN) classifier score.

Size of training set is decisive in determining the accuracy and efficiency of classifier. A small set of labeled training data may produce results with less accuracy due to high variance in the probability distribution. In [17], they tried to decrease a variance in exploiting unlabeled data by a combination of a variant of Active Learning and Expectation Maximization (EM).Our model produces better results in terms of accuracy.

NB-KNN model: Training and testing

Training of NB-KNN classifier (Class, Document)

Step 1: Vocabulary ← Extract Vocabulary from Document
Step 2: Determine the optimal value of K (for k-NN classifier)
Step 3: N ← Count Documents
Step 4: Calculate term weight according to DTE method
Step 5: Make vector of selected terms for each class
Step 6: According to optimally selected value of 'k' compute nearest neighbor s_k as: for every class calculate $p_j = |\, s_k \cap c_j \,|$ / k and keep the score p_j for each category.
Step 7: Similarly train NB classifier and calculate its score.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Step 8: for each class ∈ Category
Step 9: do $N_c \leftarrow$ Count documents in class (Document, class)
Step 10: prior[c] $\leftarrow \frac{N_c}{N}$
Step 11: $Text_c \leftarrow$ concatenate text of all docs in class (Documents, class)
Step 12: for each term ∈ V
do cond. probability[term][class]← Count tokens of term($text_c$, term)
Step 13: for each term ∈ Vocabulary
Step 14: do conditional probability [term][class] $\leftarrow \frac{T_{ct}+1}{\sum_{t'} T_{ct'}+1}$
Step 15: return Vocabulary, prior, conditional probability

Testing NB-KNN (Class, Vocabulary, prior, conditional probability, document, KNN score[class])

Step1: Weight ← Extract token from document (Vocabulary, document)
Step 2: for each c ∈ Category
Step 3: do score [class] ← log prior [class]
Step 4: for each term Weight
Step5: do score [class] += log conditional probability [term][class]
Step 6: return $\arg\min_{c \in C}$ (KNN score[Class] − NB score[Class])

VI. Experimental Results

In this section, we discuss the experimental results of the proposed methods, as compared to conventional methods. To perform the experimental task we gathered 4000 labeled financial articles from different available online news sources like Financial Express, Economics Times, Reuters, CNN Financial, CNet and Business Week.

Table 3: Accuracy measurement with different feature selection methods

# Training Documents	Bigrams	Unigrams	Weight Based Method
400	65.30%	63.34%	55.51%
800	72.45%	70.28%	61.58%
1200	74.56%	72.32%	63.38%
1600	76.24%	73.95%	64.80%
2000	79.48%	77.10%	67.56%
2400	82.34%	79.87%	69.99%
2600	80.29%	77.88%	68.25%
2800	78.79%	76.43%	66.97%

From table 3 it can be deduced that increasing size of training set have positive impact on accuracy in each category of feature selection method. Bigrams and Unigrams feature selection approach proved better result as compare to conventional weight based method. Maximum accuracy obtained at 2400 training documents with value of 82.34% and 79.87% for Bigram and Unigram feature selection method used along NBKNN respectively. For higher values of documents (i.e.>2400) accuracy decreases due to increase in dispersion and noise in feature set.

Table 4: Distribution of news articles for each class

Class	HRP	MRP	LRP	ZRP	N	Total
Train Set	404	356	459	467	714	2400
Test Set	227	263	306	131	673	1600

Table 4 describes the distribution of news articles for each class. Training and testing data are selected from randomly collected articles. Neutral class contains more data as the randomly collected financial articles have many articles which did not contain any useful information by which they can be categorized in defined categories. We aimed our experiments to evaluate the performance of classification method on the basis of feature extraction method used as compared to conventional method. The experiment was performed to show the performance of feature selection method with combined NB-KNN method. We make training and testing set out of collected 4000 articles. Training set consists 60% (i.e. 2400) articles and testing set contains remaining. As neutral class consists 30% of the collected data so we can assume that a method is able to gain 30% accuracy when it answers consistently.

Table 5: Comparison of NBKNN with other techniques

Classifier	*NB	*DT	*RC	*KNN	NBKNN
HRP	79.3	78.2	72.2	80.1	80.9
MRP	84.0	79.2	64.2	63.9	85.9
LRP	75.4	77.2	61.7	62.5	82.2
ZRP	78.6	79.0	69.3	76.9	85.2
Neutral	82.2	79.1	65.5	72.5	88.8
Average	79.9	78.5	66.6	71.2	84.6

*NB=Naive Bayes; *DT=Decision Tree; *RC=Rocchio's Classifier; *KNN=K-Nearest Neighbor (Accuracy measured in %)

Table 5 shows results of testing the accuracy performance of each classifier based on precision accuracy measure. Total 40 trials were carried out for each method. At each trial, 50 unlabeled news articles from testing set were given to each method. When less then 2400 labeled data feeds on training, the performance of the proposed method is going up until making use of 1600 unlabeled news articles, and shows the best performance on accuracy measure at the point. From this we take 2,000 news articles to make a classifier with ~85 % accuracy because it seems to largely depend on the fact that most of news providers delivered financial news with a restricted vocabulary set.

As another goal of our task is to classify the label of the financial news articles, the second experiment was performed to show the accuracy of the latest financial news data.

Table 6: Accuracy Measure of NBKNN on Testing Set

Category	Precision (%)	Recall (%)	F1-Measure (%)
HRP	84.3%	79.4%	54.7%
MRP	89.5%	84.3%	65.5%
LRP	85.6%	80.7%	57.5%
ZRP	88.8%	83.6%	64.0%
Neutral	92.5%	87.1%	72.3%
Average	88.1%	83.0%	62.6%

A news data set is made up of the articles that gathered from the same news sources as the labeled data set and reports the latest financial news at the experimental time. At each trial, 30 news articles for a company were gathered from various news sources. Table 6 shows a result NBKNN on testing set after training NBKNN with 2400 labeled data. Accuracy is measured in terms of Precision, Recall and F1-measure.The NBKNN resulted into remarkably increased performance up to 88 % with consistent classification which provides sufficient information about effectiveness of NBKNN.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

VII. Conclusion

By taking our insight to available financial news we introduced application of text classification to predict or classify the risk in investment in stock of a particular company. Text classification enormously utilizes the combined concepts of Naïve Bayes and K-Nearest Neighbor which considers refereed company's financial well being. The proposed algorithm based on co-located phrases (Bigrams) of a class predicts the category on the basis of their respective majority of weight. NBKNN outperformed as compared to other classifiers. With the proposed NBKNN technique we are able to achieve improved accuracy of 88.1%. The successful results supports that the proposed algorithms effectively works in this task, though the domain of classification problem was confined but despite of this we obtained promising result. But the proposed method has several weak points that prevent it from reaching the performance above 88 % accuracy. Failure of our approach takes place when we are having commensurate number of co-located phrases of each class as it is difficult to determine the class. To illustrate, "Stock of company X goes up as it shows profit of 15% as compare to its rival company Y which shows 5% decline in same domain". In such cases classifier will not be able to classify as the text contains commensurate co-located terms. Hence we may not be able to predict that company X is performing well and belongs to "Zero Risk Potential" because both phrases, which are "stock" with "profit" and "stock" with "decline", are very strong indicators of company's financial well-being at the moment, even though they did not indicate the same company and are not assigned with the same weight value during the training phase. Another disadvantage is that it does not concerned refereed sentence. In other words, that it does not consider sentences, which did not mention company's name as the financial evidence. To cope with these problems, we consider employing several natural language processing techniques which can provide discriminative view about misclassification. We are also projecting the approach by using discriminate term extraction with coherent semantic structure.

References

[1] A. Broder, M. Fontoura, E. Gabrilovich, A. Joshi, V. Josifovski, and T. Zhang. Robust classification of rare queries using web knowledge. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 231–238, Amsterdam, The Netherlands, 2007.

[2] A. Jain and R. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988.

[3] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes. Multinomial naive bayes for text categorization revisited. AI 2004: Advances in Artificial Intelligence, 3339:488–499, 2004

[4] A. McCallum and K. Nigam. Employing em and pool-based active learning for text classification. In Proceedings of InternationalConference onMachine Learning, pages 359–367, 1998.

[5] D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In Proceedings of the Eleventh International Conference on Machine Learning, pages 148–156, 1994.

[6] G. D. Guo, H. Wang, D. Bell, Y. X. Bi, and K. Greer. Using kNN model for automatic text categorization. Soft Computing, 10(5):423–430, 2006.

[7] P. Frasconi, G. Soda, and A. Vullo. Text categorization for multi-page documents: a hybrid naive Bayes HMM approach. In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, pages 11–20. ACM Press New York, NY, USA, 2001.

[8] S. Dumais. Using svms for text categorization. IEEE Intelligent Systems, 13(4), 1998.

[9] Porter Stemmer, http://tartarus.org/~martin/PorterStemmer

[10] R. Schapire, Y. Singer, and A. Singhal. Boosting and Rocchio applied to text filtering. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 215–223, Melbourne, Australia, 1998.

[11] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexingb by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[12] Penn Treebank Tag set http://www.ims.unistuttgart.de/projekte/CorpusWorkbench/CQP-HTMLDemo/PennTreebankTS.html

[13] S. Gao, W.Wu, C. H. Lee, and T. S. Chua. A maximal figure-of-merit (MFoM)-learning approach to robust classifier design for text categorization. ACM Transactions on Information Systems, 24(2):190–218, 2006

[14] S. Tan. An improved centroid classifier for text categorization. Expert Systems with Applications, 35(1-2):279–285, 2008.

[15] S. Weiss, C. Apte, F. Damerau, D. Johnson, F. Oles, T. Goetz, and T. Hampp. Maximizing text-mining performance. IEEE Intelligent Systems, pages 63–69, 1999.

[16] V. Lertnattee and T. Theeramunkong. Class normalization in centroid-based text categorization. Information Sciences, 176(12):1712–1738, 2006.

[17] Y. Yang and J. Pedersen. A comparative study on feature selection in text categorization. In Proceedings of International Conference on Machine Learning, pages 412–420, 1997.

[18] Z. Cataltepe and E. Aygun. An improvement of centroid-based classification algorithm for text classification. IEEE 23rd International Conference on Data Engineering Workshop, 12 :952–956, 2007.

ISBN: 978-988-17012-8-2
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)