ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

Using Training Set Selection Methods to Improve Market Prediction via News Headlines

Arti Nivrutti Korde, B.P.Kolhe

Student, Department of Computer Science and Engineering, MSS Collage of Engineering and Technology, Jalna, India

Guide, Department of Computer Science and Engineering, MSS Collage of Engineering and Technology, Jalna, India

ABSTRACT: Presently, the development of technology and social media, people's opinions and news about a particular product or currency is significantly increased. In addition to increasing volumes of text data, the unstructured characteristics of them make the analysis of these types of data with the important challenges. In this work, the propose method use different Training Set Selection (TSS) techniques like: ModelCS, ENN, MultiEdit, AllKNN, POP and MENN. Additionally, the target's features will be selected to improve the prediction of currency market (dollars, Euros) based on news headlines related to market. The evaluation of the propose method will be done by 3-layered algorithm and the experimental results will show the propose method will achieved more robust accuracies than the others.

KEYWORDS*:* Market Prediction, Text Mining, Training Set Selection

I. INTRODUCTION

1. Background

The volume of text data with the spread of technology is increasing and today a large amount of data, including people's opinions about various products, news, political, economic and other conditions are like unstructured text. The impact of news, economic conditions in the different companies and political conditions on the financial markets and the currency market is no secret, so the prediction of market changes and fluctuations based on news and other unstructured data are important for many traders and merchants. The knowledge of market forecasts using unstructured text is the same text classification that a label is given to any news data based on a set of training data with tag, which shows the direction of the market based on reports.

Generally, the market forecasting methods are divided into two methods, fundamental analysis and technical analysis. Methods that fall in the category of technical analysis use the market history data (price volatility in recent years) to predict the market and they believe that market history repeats. But fundamental analysis methods to predict in addition to historical data use unstructured text data. The propose method in this study is placed in the category of fundamental analysis methods. The main goal of this paper is to assess the influence of TSS methods on classification accuracy and forecast newsgroups. In this study, we will improve the algorithm provided by Arman Khadjeh Nassirtoussi and et al. Their algorithm predicts the direction of market volatility for the euro and dollar currencies based published headlines in the period 2008 to 2012 and its one of the best algorithms in the field of market prediction. In order to improve their algorithm, TSS methods like ENN, AllKNN, ModelCS, POP, MENN,

MultiEdit and feature selection method based on the features of the test samples are use, and the usage of these two methods in addition to increasing prediction accuracy has also significantly reduces the dimensions of feature space.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

2. Aim and Objective

The purpose of this study is to create a new market forecast field based on the headlines by using the TSS methods.

1) Using the convenient methods for selecting features such as information gain in order to use the features that are effective in the classification process not only dependent on the characteristics of the target sample.

2) Using other sampling methods to evaluate their impact on the performance of the algorithm.

3) Using two words, three words and more words features instead of using words bag selection method in pre-processing step.

II. REVIEW OF LITERATURE

1. Farzad Niknam and Aliakbar Niknafs, Kerman, Iran, **"Using Training Set Selection Methods to Improve Text Mining on Market Prediction via News Headlines"**. Second International Congress on Technology, Communication and Knowledge (ICTCK 2015) November,
11-12, 2015 - Mashhad Branch, Islamic Azad University, Mashhad, Iran
The proposed method utilized different Training Set Selection (TSS) approaches like: ModelCS, ENN, MultiEdit, AllKNN, POP and MENN. Moreover, the target's features have been selected to improve the prediction of currency market (dollars, Euros) based on news headlines. The evaluation of the proposed method has been done by 3-layered algorithm.

2. A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "**Text mining of news-headlines for FOREX market prediction: AMulti-layer Dimension Reduction Algorithm with semantics andsentiment**," *Expert Systems with Applications,* vol. 42, pp. 306- 324, 2015.

Approach is proposed to predict intraday directional-movements of a currency-pair in the foreign exchange market based on the text of breaking financial news-headlines. The motivation behind this work is twofold: First, although market-prediction through text-mining is shown to be a promising area of work in the literature, the text-mining approaches utilized in it at this stage are not much beyond basic ones as it is still an emerging field. This work is an effort to put more emphasis on the text-mining methods and tackle some specific aspects thereof that are weak in previous works.

3. Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, David Chek Ling Ngo, "**Text mining for market prediction: A systematic review**". Expert Systems with Applications, vol. 41, pp. 7653-7670, 2014.
The proposed approach review the related works that are about market prediction based on online text-mining and produce a picture of the generic components that they all have. We, furthermore, compare each system with the rest and identify their main differentiating factors.

4. E. J. de Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "**Evaluating and understanding text-based stock price prediction models**," *Information Processing & Management,* vol. 50, pp. 426-441, 2014.

This paper proposes a stock price prediction model, which extracts features from time series data and social networks for prediction of stock prices and evaluates its performance. In this research, we use the features such as numerical dynamics (frequency) of news and comments, overall sentiment analysis of news and comments, as well as technical analysis of historic price and volume. We model the stock price movements as a function of these input features and solve it as a regression problem in a Multiple Kernel Learning regression framework.

5. S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "**Combining technical analysis with sentiment analysis for stockprice prediction**," in Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on, 2011, pp. 800-807.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

This paper proposes a stock price prediction model, which extracts features from time series data and social networks for prediction of stock prices and evaluates its performance. In this research, we use the features such as numerical dynamics (frequency) of news and comments, overall sentiment analysis of news and comments, as well as technical analysis of historic price and volume. We model the stock price movements as a function of these input features and solve it as a regression problem in a Multiple Kernel Learning regression framework.

III. SYSTEM OVERVIEW/SYSTEM ARCHITECTURE

Proposed method in this paper has three main steps. In the first step of proposed method, newsgroups are entered into pre-processing stage. This stage includes stop words remove, stemming, and HYPERNYM Abstraction methods and in this stage the presence or absence of the each term in Word Net dictionary is checked. If a term does not exist in the Word Net dictionary or HYPERNYM's term doesn't be available then the term is removed from document. In the second step in order to make feature vectors by using TF-IDF and Sum Score weighting methods, appropriate weight is assigned for each term in each document. If the weight assigned to all the feature of any document is zero in order to give weight to document only the method used is TF-IDF.
In the final and third step of this method, instead of using several models, only one model is used to predict all test samples. For this purpose, at the first the terms that for any test samples are non-zero weight are selected as the superior features. Since the non-zero features of a document shows that term exist in document, so the terms with non-zero weight are more important in predicting documents. Selecting terms with non-zero weight in at least one test sample, in addition to reducing the dimensions of the feature space, provides the use of one model to predict the test samples that reduces the number of models and thus reduces the amount of computation.

Advantages of proposed approach
1. Taking the advantages of training just one model to predict the test samples and reducing the training calculation amount.
2. Using sampling methods to reduce noisy samples, which have a significant impact on increasing the performance of the algorithm.
3. Increasing the accuracy of forecasting and keeping the predictions away from the chance and also reaching the maximum accuracy in the classification of low test samples.

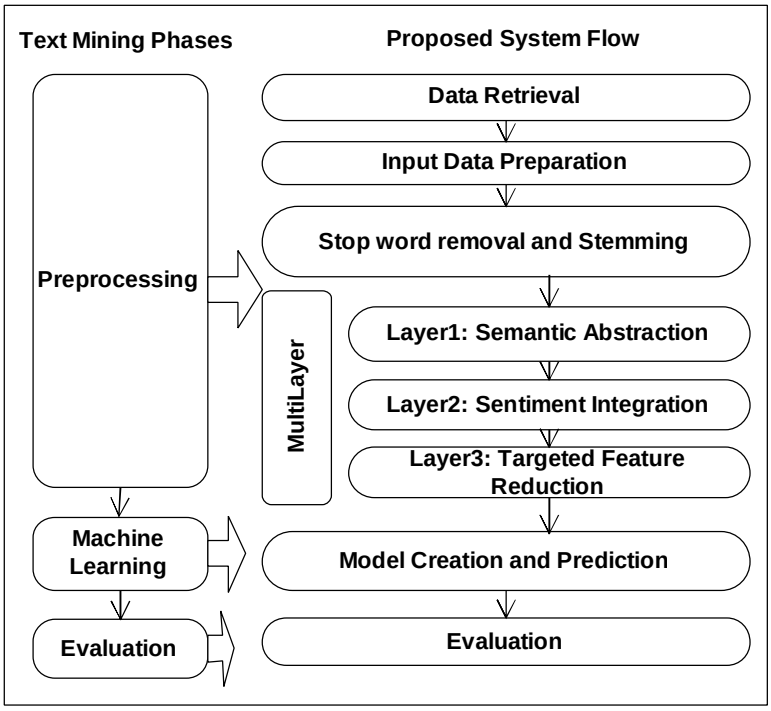

Fig.1. Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

IV. USING OTHER SAMPLING METHODS TO EVALUATE THEIR IMPACT ON THE PERFORMANCE OF THE ALGORITHM.

1. Sampling methods:

Sampling methods are classified as either *probability* or *non probability*. In probability samples, each member of the population has a known non-zero probability of being selected. Probability methods include random sampling, systematic sampling, and stratified sampling. In non probability sampling, members are selected from the population in some nonrandom manner. These include convenience sampling, judgment sampling, quota sampling, and snowball sampling. The advantage of probability sampling is that sampling error can be calculated. Sampling error is the degree to which a sample might differ from the population. When inferring to the population, results are reported plus or minus the sampling error. In non probability sampling, the degree to which the sample differs from the population remains unknown.

2. Random sampling is the purest form of probability sampling. Each member of the population has an equal and known chance of being selected. When there are very large populations, it is often difficult or impossible to identify every member of the population, so the pool of available subjects becomes biased.

3. Systematic sampling is often used instead of random sampling. It is also called an Nth name selection technique. After the required sample size has been calculated, every Nth record is selected from a list of population members. As long as the list does not contain any hidden order, this sampling method is as good as the random sampling method. Its only advantage over the random sampling technique is simplicity. Systematic sampling is frequently used to select a specified number of records from a computer file.

4. Stratified sampling is commonly used probability method that is superior to random sampling because it reduces sampling error. A stratum is a subset of the population that shares at least one common characteristic. Examples of stratums might be males and females, or managers and non-managers. The researcher first identifies the relevant stratums and their actual representation in the population. Random sampling is then used to select a sufficient number of subjects from each stratum. "Sufficient" refers to a sample size large enough for us to be reasonably confident that the stratum represents the population. Stratified sampling is often used when one or more of the stratums in the population have a low incidence relative to the other stratums.

5. Convenience sampling is used in exploratory research where the researcher is interested in getting an inexpensive approximation of the truth. As the name implies, the sample is selected because they are convenient. This non probability method is often used during preliminary research efforts to get a gross estimate of the results, without incurring the cost or time required to select a random sample.

6. Judgment sampling is a common non probability method. The researcher selects the sample based on judgment. This is usually an extension of convenience sampling. For example, a researcher may decide to draw the entire sample from one "representative" city, even though the population includes all cities. When using this method, the researcher must be confident that the chosen sample is truly representative of the entire population.

7. Quota sampling is the non probability equivalent of stratified sampling. Like stratified sampling, the researcher first identifies the stratums and their proportions as they are represented in the population. Then convenience or judgment sampling is used to select the required number of subjects from each stratum. This differs from stratified sampling, where the stratums are filled by random sampling.

8. Snowball sampling is a special non probability method used when the desired sample characteristic is rare. It may be extremely difficult or cost prohibitive to locate respondents in these situations. Snowball sampling relies on referrals from initial subjects to generate additional subjects. While this technique can dramatically lower search costs, it comes at the expense of introducing bias because the technique itself reduces the likelihood that the sample will

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

V. BAG-OF-WORDS MODEL

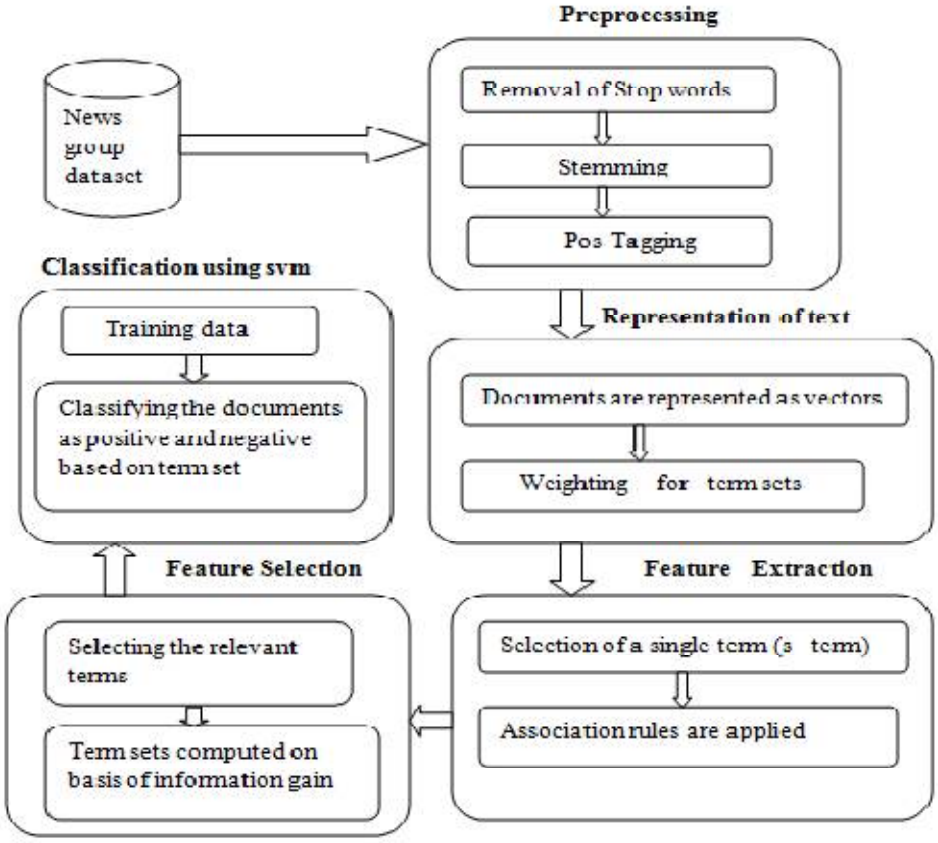

Fig No 02 BAG-OF-WORDS MODEL

VI. SYSTEM ANALYSIS

The expected results of the propose method based on different TSS approaches on classification accuracy and the reduction of training samples number. As mentioned in this section in the propose algorithm eight-way TSS (AllKNN, ENN, MultiEdit, ModelCS, POP, MENN) is, used to select the appropriate samples for training the KNN classifier that the results are visible in Table and Fig.2. As visible in TABLE II. And Fig.2 when ENN and AllKNN methods are used for sampling, the performance of the propose algorithm is better.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

No. test sample	AllKNN	ENN	ModelCS	MultiEdit	POP	MENN	3-layer algorithm
6	1	1	1	**0.8333**	1	**0.8333**	0.8333
12	0.6667	**0.9166**	0.3333	0.6667	0.3333	0.3333	0.8333
18	0.6111	**0.6667**	0.5	0.6111	0.4444	0.3333	0.6667
24	0.5833	**0.6666**	0.5	0.375	**0.625**	0.4166	0.6250
30	**0.6333**	**0.6666**	**0.6**	0.5333	**0.5667**	0.4	0.5667
36	**0.6111**	**0.5555**	**0.6111**	**0.5833**	**0.6111**	0.3888	0.5278
42	**0.5714**	**0.6666**	**0.619**	0.4523	0.5238	0.4761	0.5714
48	**0.6041**	**0.6041**	**0.625**	0.4583	0.5416	0.5	0.6042
54	**0.5555**	**0.5925**	**0.574**	0.5	0.4814	0.4814	0.5556

Table. Accuracy of propose method in comparing with 3-layer algo.

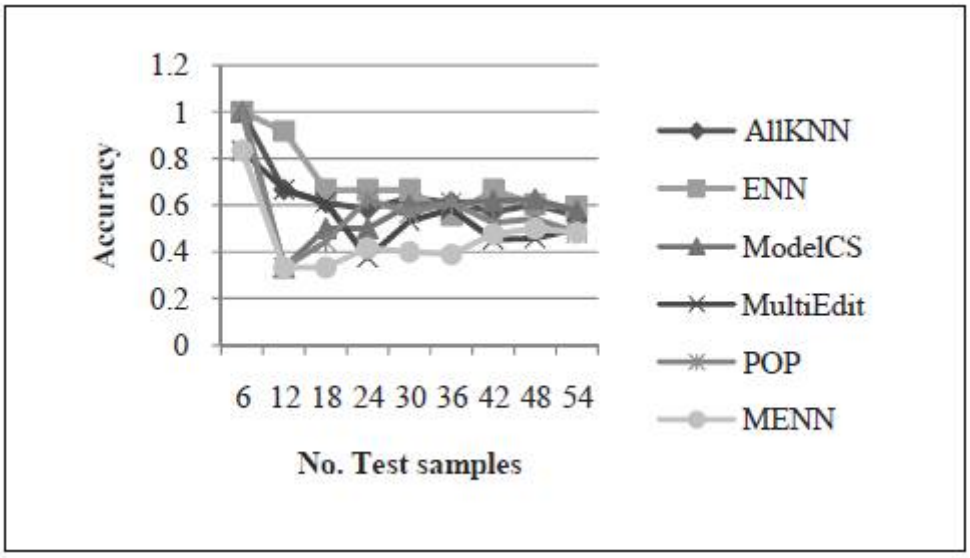

Fig.2. Compare accuracy of proposed method with four different TSS methods

V. CONCLUSION

The purpose of the propose method in this study is to show the impact of TSS methods, on performance of the proposed algorithm. For this purpose, six TSS methods (AllKNN, ENN, MultiEdit, ModelCS, POP and MEN) are used in order to eliminate noises. The propose method in the first step, applies the pre-processing techniques and semantic abstraction (HYPERNYM). Then the TF-IDF and Sum Score weighting algorithms assign appropriate weights to features vector. In the third stage, Training sample are listed by TSS methods to remove noisy data. Finally, target features will be selected.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 9, September 2017

REFERENCES

[1] Farzad Niknam and Aliakbar Niknafs, Kerman, Iran, "Using Training Set Selection Methods to Improve Text Mining on Market Prediction via News Headlines". Second International Congress on Technology, Communication and Knowledge (ICTCK 2015) November, 11-12, 2015 - Mashhad Branch, Islamic Azad University, Mashhad, Iran.
[2] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment," Expert Systems with Applications, vol. 42, pp. 306- 324, 2015.
[3] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, David Chek Ling Ngo, "Text mining for market prediction: A systematic review". Expert Systems with Applications, vol. 41, pp. 7653-7670, 2014.
[4] E. J. de Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Evaluating and understanding text-based stock price prediction models," Information Processing & Management, vol. 50, pp. 426-441, 2014.
[5] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, "Combining technical analysis with sentiment analysis for stock price prediction," in Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on, 2011, pp. 800-807.