TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 1712 - 1717

Financial News Analysis for Moroccan Stock Trend Predictions

El Bousty Hicham[1] and Krit Salah-Ddine[2]

[1]*Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Laboratory of Engineering Sciences and Energy, Agadir,Morocco*
[2]*Professor of computer sciences and Physics at Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Agadir,Morocco*

Article Info
Volume 82
Page Number: 1712 - 1717
Publication Issue:
January-February 2020

Article History
Article Received: *14 March 2019*
Revised: *27 May 2019*
Accepted: *16 October 2019*
Publication: *07 January 2020*

Abstract

This article aims to predict Moroccan stock trends based on financial news articles. Data are collected from boursenews.ma. All news collected for a single stock are lower to fit any machine learning algorithm, thus they are all combined for the training and the test issues. In these experiments we used 1061 articles published from 2015 to 2019. We compared performance of Support Vector Machine (SVM), Naïve Bayes, k Nearest Neighbors (KNN) and Decision Tree algorithms. Comparisons are performed first on news headlines and then on news corpus. Later we tried to enhance accuracy of the results using a specific dictionary-based approach. We attend 60% accuracy which is an acceptable rate in this context of the Moroccan market. We also inspected the reactivity of the Moroccan market to the publication of financial news by varying the forecasting scope.Results shows that the Moroccan Market react better to news publication four days later.

Keywords: *Machine learning, Moroccan Stock forecasting, Financial News analysis, Bag of words..*

I. INTRODUCTION

Predicting financial assets evolution retains the intention of a large part of academics (economists, mathematicians, computer scientists), practitioners and also individuals seeking to make profits by speculating on the likely prices' movement. However, and according to Efficient Market Hypothesis, no one can beat the market since price formation is a random process and already reflect all available information [11]. EMH supposes that all individuals have access to all available information, and treats individuals as rational and uniform regardless their experience, emotions, and strategies. The EMH is widely criticized, because these postulates are ideal and far from being realistic. Though, the emergence of another theory under the name of behavioral finance which claims that psychology influence financial decision-making. Indeed, Investors decision depends on their cognitive biases psychological and heuristic variables [1], all these parameters may involve an irrational behavior. Weaknesses raised about EMH explain the success of some trading strategies in particular fundamental and technical approaches. The first one is based on the analysis of available information about the company strength, the markets and the economy in general. It is interested in evaluating its real value according to a set of macro and micro economic parameters. On the other hand, the technical analysis is mainly compiling the evolution of historical prices and trading volumes to estimate values and trends. It is largely based on the graph's analysis.

Lately, machine learning algorithms are used to enhance the performance of forecasting models. Powerful algorithms are combined with technical or fundamental strategies to uncover hidden patterns. The use of machine learning doesn't only enhance the accuracy of the results but also enable the automated treatment of huge amount of data. One machine learning technique used in forecasting stocks is the text analysis technique. Text is divided to a number of tokens (word, sentence,...) and the reaction of a share to these tokens are analyzed in the purpose of building a forecasting model. This technique is usually based on the identification of sentiment behind a part of text or recognizing some fundamental analysis patterns from published reports and news. Text identified as positive may lead to excessive demands on a stock, then the value of share is likely to increase and vice versa.

This work is an attempt to build a model for forecasting Moroccans shares based on French text news. The motivation behind this work is twofold:

first, studying how reactive is the Moroccan stock market to financial text news and second, use a new model of features vectorizer using most relevant tokens retrieved from separate financial news corpus.

The rest of this work is organized is as follows. Section 2 introduces some previous research work on predicting stocks through text analysis. Section 3 presents method used for retrieving and preparing datasets. Section 4 describes the proposed model. Section 5 depicts realized experiments. Sections 6 shows the results. And finally, Section 7 concludes the contribution of this research work.

II. RELATED WORKS

Forecasting through text analysis is an active research area. The most intuitive method is the bag of word technique. In this method each unigram constitutes an entry in the feature-vector matrix. This one is formed by calculating the occurrence of each word in a corpus. Every text is represented by a line in the matrix; so, we have as many lines as corpuses to be analyzed. This technique may present high dimensionality issues and usually features should be reduced to get more accurate results [5]. First words to remove are generally stop words and common words to all entries. Lemmatization is another efficient tactic in reducing feature vector [3], hence all inflected words are regrouped in a single feature. In [8], authors used a semantic reduction process through heuristic hypernym in which each word is replaced by its hypernym from Wordnet dictionary. Some other works restricted feature vector to specific grammatical categories (for example: adjectives) determined via Part of Speech (POS) tags. Hatem Ghorbel and David Jacot used POS tags to handle negation and distinguish words that share the same spilling but different meaning (homographs) [3]. This technique seems to be very useful when trying to determine sentiment orientation. Indeed, negation forms inverse the polarity of any word. Researchers in paper [2] [7] used a dictionary approach to identify the polarity of each word in a document. Based on a collected list of negative and positive words, they determine the general orientation of a document by calculating the number of positive and negative words. Other authors used built-in dictionary for identifying emotions reflected in a document i.eSentiwordnet[3] and Linguistic Inquiry and Word Count dictionary (LIWC) [9] Authors in [5] combined all these technics to build a multi-layer features reduction's algorithm. The first layer is a semantic abstraction throw heuristic hypernym, then they use a sentiment integration layer in which a sentiment scaled score by TF-IDF(Term Frequency-Inverse Document Frequency) is attributed to each word and finally the synchronous targeted feature reduction layer that readjust the feature vector, by eliminating zeros columns, each time a new record label is to be predicted. Shaumaker and Chen [10] compared the use of Bag of Words with other technics including noun phrase and named entities. Results showed that the accuracy is close for all the three textual representations it is around 50%.

Once data is ready to be analyzed, a wide range of algorithms is available for this purpose but the most common algorithms in use are SVM [3] [5] [7] [9] [10], Naïve Bayes [5][7] [9] , KNN [5] [9]) and random forest [7].

Certainly, this area is receiving special attention from scientists. However most of the work has been applied to English. As far as French text analysis is concerned, the work consists of the translation of the texts and the use of resources available for English. Indeed, in their work of sentiment analysis of French movies reviews [3] , authors translated reviews to the English language in order to extract sentiment polarity through SentiWordnet. However, the authors are aware of the problems involved by the use of translation, in particular the disparities induced in terms of POS tags and alteration of sentiment polarity that could be observed.

III. DATA RETRIEVAL

We considered different Moroccan economic journals for collecting datasets and we compared the quantity of articles retrieved from these journals' websites. The process of extraction is described in the Figure (Fig.1). It is applied to four different journals and the results are resumed in the (Table- II) that shows the number of retrieved articles and how many of them matters in the case of this research.

Since Bourse News has most articles suitable for this research, it is retained for experimenting and evaluating our designed model. We collected 5984 articles from Bourse News. All these articles are about Moroccan economy but not all connected to a stock value. Hence, articles that are not linked to listed shares in Casablanca Stock Exchange are eliminated. Moreover, we only kept positive or negative evolution records (about 1061), this resumed the work to a two-class classification issue.

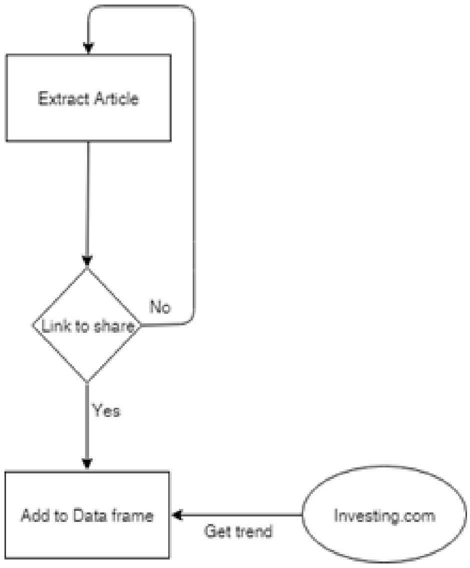

Fig.1.Datasets Extraction and Preparation

Table- I:Top recurrent shares on the BOURSE NEWS journal publications.

Share	Articles
BMCE	119
ATTIJARIWAFA	106
ALLIANCES	105
BCP	97
ADDOHA	62
CIH	53
MARSA MAROC	52
SNEP	43
SAHAM	43
COSUMAR	39
BMCI	38
DAR SAADA	31
JET CONTRACTORS	30

The most recurrent share is "BMCE" (see Table- I). However, 119 articles are not sufficient for training any machine learning algorithm. Hence, we used all these articles and try to find correlation between articles corpuses/headlines and evolution of targeted share. We extract from each article the company is talking about and get trend of this stock the day of the article publication. Our data frame looks like below (Table- III).Each row in our dataset contains date of publication, title of the news, the corpus, the concerned share and the tendency on that date (1 for up and -1 for down).

Table- II: Extracted articles

Journal	Edition language	Period	Total articles	Articles related to Moroccan shares	Retained Articles
Bourse news	French	from 23 March 2019 to 04 February 2019	5984	2770	1061
Morocco tomorrow	English	From 12 December 2011 to 24 February 2019	3302	188	110
La vie eco	French	from 04 January 2010 to 20 February 2019	11308	1143	868
Morocco world news	English	From 12 December 2011 to 11 February 2019	7388	722	479

Table- III: Clip of Data frame extracted from Bourse News

Date	Title	Corpus	Share	Evolution
04 February 2019	Financements verts : une ligne de 20 millions d'euros pour la bmci auprès de la berd	La bmci a signé le 4 février 2019, un contrat de partenariat avec la banque européenne pour la reconstruction et le développement (berd) pour le programme geffmorocco, en sa qualité de "leader dans ce segment", indique la banque dans un commu...	bmci	1
04 February 2019	Attijariwafa bank: africaine de bourse deviant "attijari securities west africa"	Africaine de bourse, société de gestion et d'intermédiation du groupe attijariwafabank dans la zone uemoa (union économiqu….	attijariwafa	1
01 February 2019	Mario camacho poursuit ses achats sur cartier saada	Cartier saada veut prospérer sur le marché américain mariocamachoinc, un des leaders de la distribution des olives de tables aux etats unis d'amérique, spécialisé dans les ve....	cartiersaada	-1
01 February 2019	Alliances : dernière ligne droite pour le reprofilage de la dette privée	Alliances vient de convoquer les porteurs de ses obligations en assemblée générale 26 février afin d'approuver le principe du remboursement des obligations et du paiem…	alliances	1

IV. PROPOSED SYSTEM

At the end of the data retrieval process, two main datasets are retained. Datasets from La Vie Eco Journal for building a financial dictionary and those from Bourse News journal are used for training and evaluating our system (Fig.2).

A. Dictionary preparation

The approach used in this article is a dictionary-based approach. Facing the lack of a Frenchfinancial dictionary, we used articles retrieved from La Vie Eco website to build a financial dictionary. We applied the bag of words technic to those articles. Indeed, the corpuses are tokenized and all stops words are removed.

We process these datasets by a Logistic Regression classifier and we then identified most influencing words, that usually impact the evolution trend of a stock. At the end of this phase, wecollected about 400 words and all inflected forms of those words have been listed for the purpose of applying the lemmatization process. The objective of this tactic is to construct a non-biased dictionary as we use a sperate data source for training and testing our model.

B. Input data preparation

Articles collected from Bourse News in data retrieval stage is analysed in this phase. The vector matrix is formed and then handled by machine learning algorithm. Operations executed during this phase are described as follows:

- Text tokenization

In order to apply any natural language processing technic, text is generally split into tokens (words in our case), then POS tag, named entities or any other technic can be easily applied. The main goal of this operation is eliminating punctuation and storing separately each word.

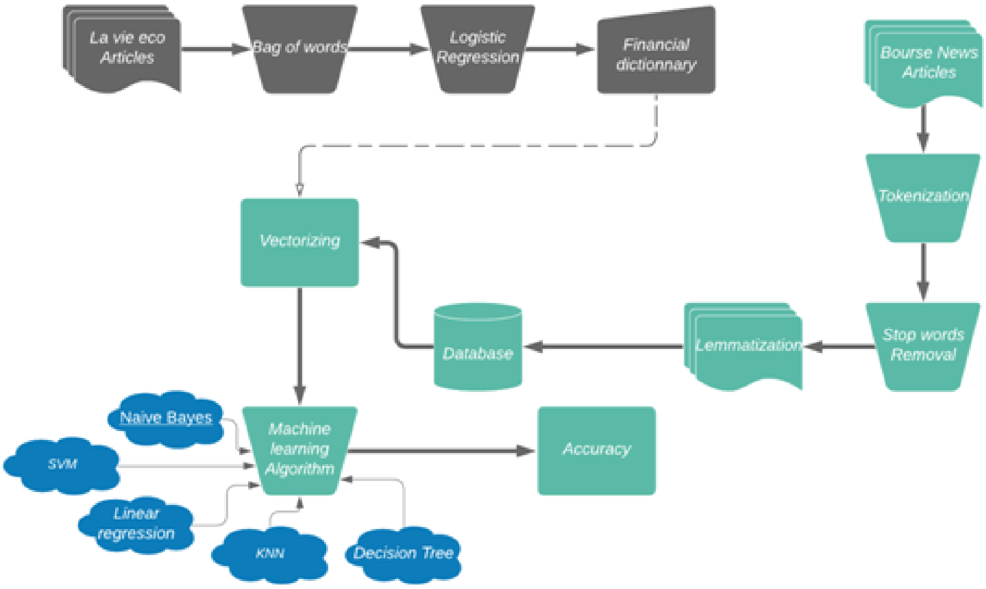

Figure 2: Prediction From Financial News Model

- Stop word Removal

The pre-processing starts with removing stop words from the corpus. Numbers, words composed from few characters (less than three characters) and thosepresent on most articles are all removed from tokens list. This step is most useful when applying a Bag of words technic in order to reduce the features space. In our case this helps optimizing time and resources for lemmatization process.

- Lemmatization

Each word in tokens list is replaced by its lemma, so all inflected forms of a word is represented by only one feature.

- Vectorization

Only words contained in the build financial dictionary are kept in the tokens list, we calculate then the occurrence of each word in the tokens list. Each article corresponds to an occurrence vector and the matrix is made by regrouping all these vectors.

V. EXPERIMENTS

For our experiments we used different machine learning algorithms in order to answer our research questions and we picked some of the best-known algorithms for text analyzing (SVM, KNN, Naïve Bayes, Decision Tree and Linear regression). Articles gathered from Bourse News website served for training and testing these algorithms.

We first started by comparing the performance of headlines-based approach and corpuses-based approach. Articles and headlines are vectorized via a traditional bag of words technic and the accuracy is reported in order to determine whether few words can predict stock movements better than the whole article.

The second experiment trained our proposed system. The financial dictionary built through the analysis of La vie eco articles is used for the purpose of vectorizing Bourse News articles.

Finally, we varied the prediction scope for the purpose of identifying how elastic is the Moroccan market to financial news and where can be placed towards the efficiency theory. We forecast shares movement 4 days and 8 days after date of article publication.

For all experiments and since we have only few datasets for training this model, we used the cross-validation approach. So, we partitioned data to five folds and each time we use four folds for training and the fifth for tests.

VI. EXPERIMENTAL RESULTS

Considering the accuracy measurement of stock predictions through financial headlines and financial articles we conclude that headlines-based approach performs better, it attends about 59% for the Logistic regression and the Decision Tree algorithms (Table IV), this is 9% better than the accuracy classification by chance when 2 classes available presumed at 50%. Feature vector for articles analysis is more complex and present high dimensionality since headlines resumed contents to few words, hence the performance drops when analyzing articles contents. The second experiment focused on evaluating our designed model. Accuracy is enhanced for all machine learning algorithms.

Improvement reaches 7% for Naïve Bayes and KNN has the best performance even better than accuracy made by Headline approach (Table V). We limited ourselves to apply our model to articles contents otherwise, the majority of the values in the feature matrix will be zeros. Experiments showed also that our model predicts more accurately stocks direction when the scope is extended to 4 days (Table VI) which allows an additional opportunity for investment. Hence, we can understand that the Moroccan market is reacting slowly to the publication of financial news contrary to what is presumed by the EMH theory.

VII. CONCLUSION AND FUTURE WORK

This work is a forecasting intraday directional movement from financial news articles. This is a unique research as it is the only paper predicting Moroccan stock market through financial news, at least to our knowledge. The accuracy in the majority of similar works is reported in the range of 50% to 70% [5], similarly our model that reached a 60% accuracy and which advance the model proposed by [10]. This performance can be enhanced if all articles correspond to a single company as demonstrated by [4]. Furthermore, most works dealing with French texts for classification or regression issues opted for translation to benefit from a wide choice of resources available for the English language. As a contribution of this work, we approached the analysis of a French corpus in a different way by building a financial dictionary from a separate resource. This method showed better performance compared to Bag of words technique. However, the built dictionary contains only 400 words. As a research direction this dictionary can be enriched with more influencing words from the finance domain. Besides, all these dictionary entries' can be sentimentally labelled; thus, a sentimental approach can be evaluated in a subsequent work.

Table-IV: Articles and Headlines predictions accuracy (bag of words technic)

	SVC	**Logistic regression**	**Naives Bayes**	**KNN**	**Decision Tree**
Headline	57,39%	59,1%	50,14%	57,86%	59,09%
Corpus	57,77%	55,61%	48,26%	57,77%	54%

Table- V: Bags of words and Our model accuracy comparison

	SVC	Logistic regression	Naives Bayes	KNN	Decision Tree
Bag of Words	57,77%	55,61%	48,26%	57,77%	54%
Our Model	58,97%	56,92%	55,21%	59,23%	57,69%

Table- VI: Accuracy according to forecasts scope.

	SVC	**Logistic regression**	**Naives Bayes**	**KNN**	**Decision Tree**
Same day	58,97%	56,92%	55,21%	59,23%	57,69%
4 days	59,11%	56,37%	54,11%	60,99%	58,54%
8 days	59,69%	56,11%	52,44%	60,70%	60,88%

REFERENCES

[1] E. Gupta, P. Preetibedi, and P. mlakra, "Efficient Market Hypothesis V/S Behavioural Finance," *IOSR J. Bus. Manag.*, vol. 16, no. 4, pp. 56–60, 2014.

[2] A. Nayak, M. M. M. Pai, and R. M. Pai, "Prediction Models for Indian Stock Market," *Procedia Comput. Sci.*, vol. 89, pp. 441–449, 2016.

[3] H. Ghorbel, "Advances in Distributed Agent-Based Retrieval Tools," vol. 361, no. May, 2011.

[4] K.-G. Aase and P. Öztürk, "Text Mining of News Articles for Stock Price Predictions," *Dep. Comput. Inf. Sci.*, vol. Msc., no. June, p. 82, 2011.

[5] A. KhadjehNassirtoussi, S. Aghabozorgi, T. Ying 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 Syst. Appl.*, vol. 42, no. 1, pp. 306–324, 2015.

[6] M. T. Hallissey, A. J. Jewkes, J. A. Dunn, L. Ward, and J. W. L. Fielding, "Resection-line involvement in gastric cancer: A continuing problem,"*Br. J. Surg.*, vol. 80, no. 11, pp. 1418–1420, 1993.

[7] K. Joshi, "S TOCK TREND PREDICTION USING NEWS SENTIMENT ANALYSIS."

[8] R. Vijayan and M. A. Potey, "Improved Accuracy of FOREX Intraday Trend Prediction through Text Mining of News Headlines using J48," vol. 5, no. 6, 2016.

[9] D. K. Kirange, J. T. Mahajan, M. D. K. Kirange, and R. R. Deshmukh, "Sentiment Analysis of News Headlines for Stock Price Prediction FIST-2013 : GIS and Remote Sensing Theme View project Speech Recognition System View project Sentiment Analysis of News Headlines for Stock Price Prediction," An Int. J. Adv. Comput. Technol., vol. 5, no. 3, pp. 1–6, 2016.

[10] R. P. Shaumaker, and H. Chen (2006) 'Textual analysis of stock market prediction using financial news articles', 2006

[11] E. Fama.(1964) 'The behavior of stock Market Prices'. Graduate School of business University of Chicago

AUTHORS PROFILE

Hicham el bousty is currently a PhD student at the Laboratory of Engineering Sciences and Energy, Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University, Morocco. He obtained his engineering degree in computer science in 2009, from the National School of Mineral Industry, Rabat, Morocco. Business intelligence and analytics are the main fields of his research. In addition to his academic career, he is responsible of the security of the Moroccan identification system.

Salah-ddineKrit is currently an Associate Professor at the Polydisciplinary Faculty of Ouarzazate, Ibn Zohr University Agadir morocco, Dr. Krit is currently The Director of Engineering Science and Energies Laboratory and The Chief of Department of Mathematics, Informatics and Management. Dr. Krit received the Ph.D degrees in Software Engineering from Sidi Mohammed Ben Abdellah university, Fez, Morroco in 2004 and 2009, respectively. During 2002-2008, he worked as an engineer Team leader in audio and power management Integrated Circuits (ICs) Research, Design, simulation and layout of analog and digital blocks dedicated for mobile phone and satellite communication systems using Cadence, Eldo, Orcad, VHDL-AMS technology. Dr. Krit authored/co-authored over 130 Journal articles.