National Conference on Innovations in Computing and Networks (NCICN) - 2019
© 2019 IJSRSET | Volume 5 | Issue 9 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099

An Automated System for Analysing the Financial News

Thanuja A[*1], Aswathy L C[*2] Ameena A R[*2], Nebila Nizam N[*2]

[*1]Asst.Professor, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Pallimukku, Vadakkevila, Kollam, Kerala, India

[*2]B.Tech Students, Department of Computer Science and Engineering, Younus College of Engineering and Technology, Pallimukku, Vadakkevila, Kollam, Kerala, India

ABSTRACT

The 24-hour news cycle and barrage of online media is a constant drum beat. The flow of positive and negative financial news is always in flux, influencing our current perspective and reassessing our future outlook. Nowhere is this more true than in the capital markets where assets are priced and risk assessed based on future expectations. While many factors influence a trader's decision to buy or sell an asset it can be argued that the sentiment from the 24-hour news cycle greatly impacts their outlook on the future value of an asset. In this paper our method propose new methods to predict the positive or negative sentiment of financial news. Using Natural Language Processing methods, our method extract syntactic sentence patterns from financial news. From these patterns we conduct experiments using machine learning sentiment analysis approaches to predict sentiment. It find that our sentiment prediction methods are able to consistently out perform the methods. Our robust techniques give the financial practitioner a method to analyze the news sentiment factor and labeling them using a machine learning algorithm logistic regression.

Keywords: Web scraping, Sentiment Analysis, Labeling, Logistic Regression.

I. INTRODUCTION

One of the pillars of modern financial theory is the Efficient Market Hypothesis (EMH). EMH was first proposed by Professor Eugene Fama of the University of Chicago in the 1960's. According to EMH, market prices reflect a security's fundamental value over the long run. However in the short run a security's price may deviate significantly from its true fundamental value. EMH further postulates that the price of securities change as new information enters the market and rational actors adjust their price expectations in reaction to this new information.

The goal of this project is to implement some fairly advanced machine learning algorithms in order to detect the overall meaning of financial news without having to read them. The project is structured in different parts. The first part is the web scraping. Using different tools like Tor it is possible to download automatically thousands of articles from different websites like Bloomberg or CNN and to archive them in different folders corresponding to different companies. The next step is labelling. This operation is needed to calibrate the dataset in order to implement correctly the learning part. This is not a computed

operation, it must be done by humans so that the overall sense of each article is understood and the data set is divided into four categories, they are: positive, negative, neutral and irrelevant. Then articles are parsed so that all the useless words are deleted and only nouns, verbs, adjectives survive, making the dataset analysis easier. This is when the learning phase begins.

The algorithm analyses the articles and builds the dictionaries with the words it found in the training set. The test set compares the words contained in the new articles with the labelled ones and assigns a label to the new article. Another technique used is the logistic regression which is more accurate because it gives more importance to more frequent words. The environment used in this project is Python which is a versatile language and has a lot of libraries already available to perform natural language processing. The possibility of simply classifying articles regarding a specific company is useful especially for possible investment strategies which are based on sentiment analysis.

Sentiment analysis is the task of extracting a person's opinion or emotional response from an object or event. Our method, compare use machine learning technique, to which we contribute. Using tools from Natural Language Processing we decompose financial news headlines to extract and build features from syntactic sentence patterns. We identify recurring patterns which are used to train our models. Lastly we conduct experiments using both methods to predict financial news sentiment and discuss the results.

The possibility of simply classifying articles regarding a specific company is useful especially for possible investment strategies which are based on sentiment analysis. The machine learning approach to sentiment analysis is a supervised classification task which involves building classifiers from labeled instances of texts or sentences.

II. RELATED WORKS

Technical analysis depends on historical and time-series data. These strategists believe that market timing is critical and opportunities can be found through the careful averaging of historical price and volume movements and comparing them against current prices. Technicians also believe that there are certain high/low psychological price barriers such as support and resistance levels where opportunities may exist. They further reason that price movements are not totally random, however, technical analysis is considered to be more of an art form rather than a science and is subject to interpretation.

While many researchers use pre-defined sentiment dictionaries such as the Harvard Inquirer (HI) or SentiWordNet argues that general lexicons are not suited to domain specific applications. They find that 73.8% of the negative words in the HI are not considered negative when used in the context of finance. To address this shortcoming they propose a new sentiment lexicon tailored to the domain of finance. Using a financial lexicon they find a significant reduction in sentiment classification errors.
Attempts to predict positive or negative market volatility in the financial markets. Taking a machine learning approach they follow. Using the Federal Reserve meeting minutes as input they select lists of words which imply market volatility. The word list is then ranked and sorted. These words become the "seeds" to generate dependency trees used in the feature vectors for model learning and testing. They find that using a considerably smaller subset of text containing only financial domain words achieves comparable results to that of using the entire corpus, also experiments with a specific lexicon constructed with words related to financial risk, such as forbear, default, etc. Using this tailored lexicon they build financial risk models to predict company specific risk. They find a strong correlation between risk words and the risk of companies. Combines financial news articles with social media content. This data is binned into fourteen categories which evoke specific emotions such as fear, joy, greed, optimism, etc. Using the binned data

they perform linear regression and use neural networks to predict FX rates. They find that the neural networks achieve superior results over linear regression methods.

While a consensus appears to be forming on the need for domain specific lexicons and training texts, there appears to be a gap in viewing these domain specific words in the context in which they appear.

The lexicon approach to sentiment analysis requires a dictionary of words and the associated sentiment score for each word in the dictionary. Sentiment word scoring typically entails a binary score of +1 for positive sentiment and -1 for negative sentiment. Neutral sentiment that is words that do not convey positive or negative sentiment such as the word "the", commonly referred to as stop words are often discarded. Analysis however requires the decomposition and examination of sentence structure so we have chosen not to discard any text and will assign a sentiment score of 0 to neutral words.

To determine the sentiment score for a sentence a Bag-Of-Words (BOW) model is followed. A BOW model considers each word to be independent. Thus each word is independently scored by looking up the word in the lexicon and assigned the associated word sentiment score. This process is repeated for each word in the sentence. The sentiment scores are then summed and the sign of the result is the final sentiment score for the text. A sentence with a greater number of positive sentiment words than negative will be scored as positive and vice versa. Popular lexicons include the Harvard General Inquirer Lexicon (H4N) developed, the MPQA Subjectivity Lexicon and SentiWordNet.

The BOW model using the H4N lexicon to assign sentence sentiment will be used to determine our baseline score in the experiments section. The H4N lexicon consists of a total of 11,789 entries (words) of which 1,917 are positive words, 2,291 negative words and the remainder is considered neutral. Performance of all other models in the study will be reported relative to the baseline score. Further details of lexicon based sentiment analysis approach may be found and relatively in more recent work.

While lexicon sentiment methods have proven useful there are several draw backs. These include the lack of domain specificity, the independence assumption, the laborious nature of building a lexicon, the absence of context and their non-robust nature due to missing words. Below the following text are the sentiment scores for each word as they are found in the Harvard General Inquirer lexicon using a BOW sentiment scoring model.

"Google beats earnings estimates by a large margin."

Not Find -1 0 0 0 0 0 Not Find

Summing the word sentiment scores yields a result of -1, taking the sign gives the sentence sentiment of negative though it is evident that the sentence emotes positive sentiment when viewed from the perspective of an investor. This highlights several of the draw backs previously noted such as lack of domain specificity, the independence assumption, lack of context and missing words. The entire sentiment score is predicated on the single word "beat" which the H4N lexicon labels as negative. The word "beat" is a homograph (it has multiple meanings), the context in which is used allows us to deduce the intended meaning. While one definition for the word "beat" is "to strike violently" it becomes obvious by the collocated word "earnings" combined with viewing through the domain of investments that "beat" in this context does not mean 'to strike violently". The intended meaning in this context is "to surpass or exceed" which should emote positive sentiment from an investor. Had the sentiment model included investments domain knowledge and was contextually aware the word "beat" would have been labeled positive and the entire sentence sentiment would flip from negative to positive.

III. MATERIALS AND METHODOLOGY

Dataset

Our articles are taken from various financial news available websites like CNN, Bloomberg etc. Each news

is annotated with its category, subcategory, dates and set of tags that describing the contents of the news. For this method our method focuses on financially related news.

Methodology

In our proposed system our method have proposed some fairly advanced machine learning algorithms in order to detect financial news without having to read them. This automated system for the analysis of news is useful and accessible for any company. System consists of 3 steps - first part is the web scraping. Using different tools like Tor it is possible to download automatically thousands of articles from different websites like Bloomberg or CNN and to archive them in different folders corresponding to different companies. The next step is labelling.

This operation is needed to calibrate the dataset in order to implement correctly the learning part. This is not a computed operation, it must be done by humans so that the overall sense of each article is understood and the data set is divided in four categories: positive, negative, neutral and irrelevant. Then articles are parsed so that all the useless words are deleted and only nouns, verbs, adjectives survive, making the dataset analysis easier.

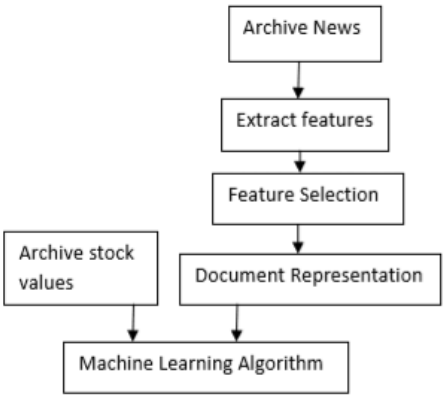


Figure 1: Work flow diagram for proposed system

Following convention positive sentiment news headlines were labeled as positive and negative sentiment headlines were labeled as negative. News headlines which did not emote sentiment were evaluated as neutral.

A. Web Scraping

Web Scraping also termed Screen Scraping, Web Data Extraction, Web Harvesting etc, is a technique employed to extract large amounts of data from websites where by the data is extracted and saved to a local file in your computer or to a database in table (spreadsheet) format.

B. Sentiment Analysis

Sentiment Analysis is the process of `computationally' determining whether a piece of writing is positive, negative or neutral. It's also known as opinion mining, deriving the opinion or attitude of a speaker. Here we analyses whether the financial news is positive, negative, or neutral.

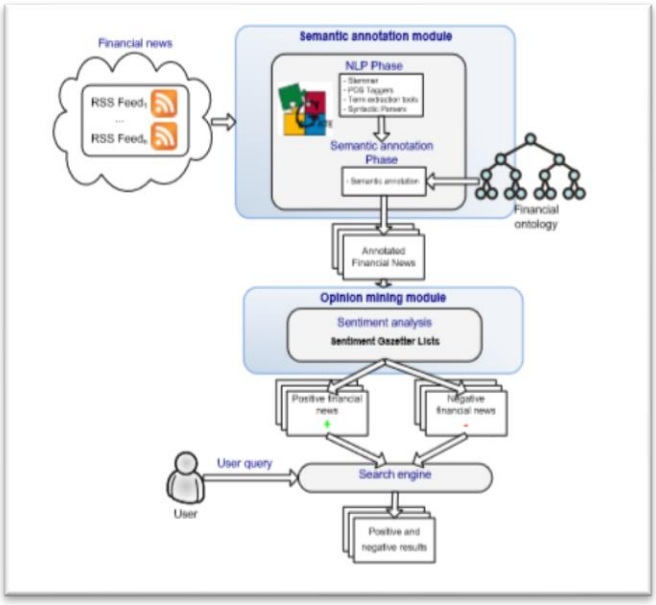


Figure 2: Architecture of the system

C. Labeling

Take a set of unlabeled data and augments each piece of that unlabeled data with meaningful tags that are informative. Using labeling we categorize the financial news as positive, negative, and neutral.

The implementation of machine learning algorithms requires labeled training data as input. One of the challenges of developing sentiment models for the financial domain is the lack of training data as input. For other domains such as discussion forums, blogs and product reviews there are a number of datasets available, but there is a lack of annotated data for financial sentiment analysis.

This study assumes that a financial news Sentiment accurately reflects the sentiment of the financial news article, thus our exercise it to label news headlines. The overall sentiment of the news article was considered, that is each word in the sentence is interpreted in the context of the entire sentence, words are not independent.

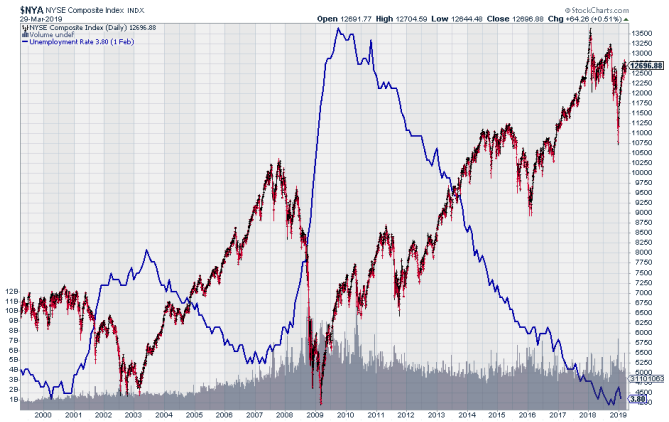
Feature Extraction

Reports the efficacy of syntactic patterns in sentiment aspect extraction. They find that the most productive pattern looks for noun sequences which follow an adjective. Generally, information about sentiment is conveyed by adjectives or more specifically by certain combinations of adjectives with other parts of speech. The adjective-noun pair method used in sentiment studies by online product review researchers has found its way into financial news analytics research in such works. While research has show the efficacy of the adjective - noun sentiment detection method for document level analysis it will not always suffice or may even break down at a more granular level. When analyzing a large document it would be highly improbable that a single adjective-noun pair would not exist, however once the analysis is narrowed to the sentence level the adjective-noun pair maybe absent, thereby proving ineffective.

IV. RESULT AND DISCUSSION

This method mainly consists of three steps. Web scraping consists in the extraction of articles from different websites, after processing scraping the first dataset is ready to be labeled and parsed. The labeling step is necessary in order to calibrate the learning algorithm. For each article in the database a label is chosen: positive, negative, and neutral based on what is written about the society in the article. In parsing, a lexical analysis using NLTK (Natural Language Tool Kit) is performed looking for the more useful lexical categories. Useless words, special characters, and punctuations are also removed from the text. Learning consists in the extraction of list a from the training set of meaningful words with a predicted positive, negative, or neutral effect is done. At the end the words will be included into a dictionary that will be used in the testing phase. After the realization of these steps each article will be labeled following the percentage of positive or negative words that are contained in the article.

s



IV. CONCLUSION

It is a system that analyses news about a company and tells if it is positive or negative news. With machine learning, based on natural language processing algorithms. Nowadays the newspaper plays an important role regarding the image of a company. With this tool, keeping track of the news without reading the articles will be possible. This automated system for the analysis of news is useful and accessible for any company.

V. ACKNOWLEDGMENT

We are sincerely thankful to our Principal Dr. P Sreeraj, for providing us the facilities in order to go ahead in the

development of our research. We express our deep and sincere gratitude to Dr Nijil Raj N, Head of Computer Science and Engineering department, Prof. Yasir .A, Project coordinator and our guide Asst.Prof Thanuja A, for providing valuable advice and timely instructions. We would like to express our very great appreciation to Mr Sreedarsh S for his valuable and constructive suggestions during the planning and development of this research work.

V. REFERENCES

[1] Baccianella, S., Esuli, A. and Sebastiani F. (2010)SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, In Proceedings of the Seventh Conference on International Language Resources and Evaluation

[2] Blitzer J., Dredze, M. and Pereira, F. (2007) Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification, in Association for Computational Linguistics

[3] Des, S. and Chen, M., (2007) Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, in the Journal of Management Science, vol 53

[4] Blair-Goldensohn, S., Neylon, T., Hannan, K., Reis, G., McDonald, R., and Reynar, J. (2008) Building a sentiment summarizer for local service reviews, In the Proceeedings of NLP in the Information Explosion Era 2008