2018 International Conference On Business Innovation (ICOBI), 25-26 August 2018, NSBM, Colombo, Sri Lanka

Analysis of Financial News Using Natural Language Processing and Artificial Intelligence

Aditya Khant
Harvey Mudd College
Claremont , California, USA 91711
Email: akhant@hmc.edu

Mahendra Mehta
NeuralTechSoft
Goregaon, Mumbai, India 400063
Email: mahendra@neuraltechsoft.com

***Abstract*—The task of analysis of financial news uses state-of-the-art Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques. The paper demonstrates how NLP techniques can be used to provide succinct summaries, identify keywords and determine sentiment of a certain news article. These features can then be used to make informed decisions that are based on recent, relevant reports. This paper also exhibits a methodology of identifying impact investment using a Naive Bayes Classifier, which can be extended to other financial terminology. We find that the usage of NLP and AI techniques such as sentiment analysis and keyword extraction enhance the information presented in online news articles by filtering out irrelevant content.**

I. Introduction

Financial markets are volatile and real time updates and analysis are of utmost importance when dealing with them. These markets are susceptible to the global events and phenomena such as trade wars, civil unrests, innovation, and scientific discoveries. Financial News is obtainable from a multitude of sources, both online and offline. Online sources here are defined as those which can be acquired via the internet and offline sources here are those which are propagated via other media. Offline sources include news and insights found via newspaper and television. News obtained via a newspaper is obsolete for a financial market as sensitive as a stock market. News found on television is live, but such news cannot be analyzed with ease. Online sources are more superior in terms of their offline, when it comes to relevance and ease of analysis.

There are many websites that distribute publish and aggregate such news. Some of them include BusinessInsider, BusinessToday and Economic Times. Each of these numerous websites have many news articles each with thousands of words which can be difficult to process for humans. Hence there is a need for aggregating and analyzing such news using modern computer science techniques which will enable humans to make informed decisions.

Such computer science techniques include Natural Language Processing (NLP) and Artificial Intelligence (AI). The definition of Natural Language Processing is as follows. It is a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications [1]. There are many helpful natural language tasks like analyzing sentiments, creating summaries and extracting keywords. Artificial Intelligence is a vast field that aims to understand the intelligent entities and enable machines to mimic these abilities [2]. These state-of-the-art technologies can provide insights into financial news.

Techniques like AZFin leverage textual analysis using NLP to predict stock market prices [3]. However, such techniques are limited and narrowly focused on singular goals and have no means of utilizing real-time data. Data Mining is another aspect that provides good insight into financial news [4]. However, such data mining occasionally results in a lot of data being displayed to the user, with little or no analysis. Our work simplifies the process of aggregating data and then uses modern NLP and AI techniques such as keyword extraction, summary generation and sentiment analysis to extract relevant financial information from news and display it to the user in a succinct manner.

The remainder of this paper is organized as follows. Section II presents the existing technological advancements in the fields of AI and NLP which form the basis for our work. Section III elaborates on how we leverage the technologies mentioned in Section II for financial news analysis. Section IV concludes and mentions directions to future work.

II. Background

Our work relies heavily on certain Natural Language Processing and Artificial Intelligence Techniques. These techniques include summary and keyword extraction using the TextRank algorithm, sentiment analysis and Naive Bayes Classification [5] [6] [7].

A. TextRank: Summary and keyword extraction

There are two state-of-the-art methods of creating summaries: extraction and abstraction. Extraction algorithms, as the name suggests, summarize an article by scoring the sentences with relevance and selecting the most relevant sentence

based on this score [8]. Abstraction algorithms utilize neural networks to gain an understanding of the meaning of the article, which is then passed on to a natural language generation program that creates a summary [9]. We prefer the extraction strategy because it is not computationally expensive.

TextRank is an unsupervised algorithm for the automated summarization and keyword extraction of texts. The algorithm applies a variation of PageRank [10] over a graph constructed of sentences in a single document. This produces a ranking of the elements in the graph according to which the most important elements are the ones that better describe the text. This approach enables TextRank to build summaries without a training corpus and makes the algorithm language independent.

The function of the original algorithm can be written as: Given S_i, S_j two sentences represented by a set of n words that in S_i are represented as $S_i = w_1^i, w_2^i, ..., w_n^i$. The similarity function for S_i, S_j can be defined as:

$$Sim(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{log(|S_i|) + log(|S_j|)} \quad (1)$$

The result of this algorithm is a dense graph representing the document. From this graph, PageRank is used to compute the importance of each vertex. The most significant sentences and words are selected to form the summary and keywords [11].

B. Sentiment Analysis

The process of analyzing sentiment relies on training the machine to identify positive and negative words using a Word Net . Each word is scored using a p normalized Kendall τ score. The function to calculate this score is formalized as follows:

$$\tau_p = \frac{n_d + p \cdot n_u}{Z} \quad (2)$$

where, n_d is the number of discordant pairs, i.e., pairs of objects ordered one way in the gold standard and the other way in the tested ranking; n_u is the number of pairs which are ordered (i.e., not tied) in the gold standard and are tied in the tested ranking; p is a penalty that is attributed to each such pair, set to $p = \frac{1}{2}$ (i.e., equal to the probability that a ranking algorithm correctly orders the pair by random guessing); and Z is a normalization factor (equal to the number of pairs that are ordered in the gold standard) [12].

C. Naive Bayes Classification

Bayes Classifiers rely on Bayes Theorem. Bayes Theorem is formalized by:

$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)} \quad (3)$$

where $p(A|B)$ is the probability of A given the probability of B, $p(B|A)$ is the probability of B given the probability of A, $p(A)$ is the probability of A, and $p(B)$ is the probability of B [13].

The classifier algorithm requires a labeled text dataset.The algorithm first identifies various textual features in the text dataset. It works out the probability of finding these features for every unique label in the dataset. These probabilities make up the model of the classifier. When a new unlabeled dataset is given as input to the algorithm, it computes probability of the features found in the model and multiplies all of them. It then compares it to the model and labels the text.

III. News Aggregation and Analysis

Our work comprises of the following stages: aggregating news, preprocessing the raw data, extracting the summary and keywords from the source text, analyzing the sentiment of the summary and classifying the text as an impact investment.

A. News Aggregation

For the purpose of aggregating news, we use sources found on the Internet, that is, online news sources instead of offline news sources like newspaper and television. This is done to automate the process of news aggregation and get news updates in real-time. The process of aggregating news is automated using a combination of using web scraping to crawl the Internet for news articles and the Bing Search Engine's News API [14]. The reason we utilize both the sources in conjunction is to increase the categories and articles of news found and to minimize duplication.

The advantages of crawling websites that published financial news was that they sorted their information by sector in which the news was involved. This provides a more focused and relevant output to the user. The Bing Web Search API is used to find articles for user-specified queries that are not a part of the pre-determined sectors on financial news websites. Both tools returned data in the raw HyperText Markup Language (HTML) format. The aggregated data needed to be preprocessed for analysis.

B. Data Preprocessing

Data preprocessing is important for accurate summarization and keyword extraction. The data that is received using the aggregation is in a raw HTML state. This means that it has unwanted information like HTML tags and advertisements which will result in inaccurate summaries. The preprocessor first identifies all the paragraphs in the raw HTML file and extracts text from them while discarding other tags such as buttons and images. It also identifies the title from the heading tag in the raw. It tries to identify the date published by looking for words such as "date", and "published on". The title, clean text and date are then ready for summarization and keyword extraction.

C. Summarization and Keyword Extraction

To create a summary and extract keywords we use the methods described in Section II A which have been implemented in the summaNLP python library [11]. The ratio of the number of words in the summary to the original article is 1:5. We find that this ratio creates a succinct summary while preventing

excessive loss of information. This summary is then displayed to the user. Sometimes the articles scraped are really small and cannot be summarized into a meaningful summary. This is usually when the summary is fewer than 4 sentences long. When this occurs, the algorithm return the original article of the text. For keyword extraction, the source text is processed using summaNLP. Our algorithm then identifies duplicates in the keywords output such as plurals, verb/noun forms and different tenses by identifying the stem words of each of the keywords. The duplicates are eliminated and the remaining keywords are displayed to the user.

D. Sentiment Analysis

After extracting the summary, we analyze its sentiment by finding out its polarity. Polarity is a measure of the positivity of the article. We do it using an implementation of the steps mentioned in Section II B. This polarity analysis ranges from -1 to +1 where -1 is negative sentiment, 0 is neutral and +1 is positive sentiment. This provides an easily understandable measure for the user and reduces the time spent on learning about the financial state of a certain institution or sector.

E. Impact Investment Analysis

Impact Investment is investment in businesses that prioritize social and environmental impact over profit. They differ from non-profit organization and Non Governmental Organizations (NGOs) because impactful businesses operate for a profit [15]. To analyze whether a business is impact investment or not, given an article about it, we use the Naive Bayes Classifier Method mentioned in Section II C. The process of training a machine to identify impact investment businesses involves the following steps: Data gathering, data preprocessing, model building and accuracy testing.

1) Data Gathering: We created a labeled dataset of positive and negative articles about businesses in a variety of sectors, where positive articles were those which contained impact investment and negative articles were those which did not contain impact investment business. These articles were aggregated using the web scraping process mentioned in Section III A.

2) Data Preprocessing: This data required more preprocessing than the one mentioned in Section III B as the Naive Bayes Classifier (NBC) relies on relationships between words and is case sensitive. The HTML tags and non-letter punctuations are removed from the raw text. This is done to avoid the influence of periods and commas on the result of our model. The entire text is converted to lowercase as the NBC is case sensitive while learning. Stop words, words which are irrelevant to the topic like article, are removed from the text. This improves the models accuracy as it learns to use only relevant textual data [16].

3) Model Building: The clean preprocessed data is used to build the Impact Investment Analysis model using the NBC method mentioned in section II C. Our model used eighty percent of the original data for training purposes. Our model had the following most important features:

contains(online) = True neg : pos = 5.8 : 1.0
contains(time) = True neg : pos = 5.1 : 1.0
contains(tools) = True neg : pos = 4.0 : 1.0
contains(end) = True neg : pos = 4.0 : 1.0
contains(industry) = True neg : pos = 3.8 : 1.0
contains(people) = True pos : neg = 3.2 : 1.0
contains(like) = True neg : pos = 3.1 : 1.0
contains(income) = True pos : neg = 2.9 : 1.0
contains(services) = True pos : neg = 2.9 : 1.0
contains(tech) = True pos : neg = 2.9 : 1.0

The above out demonstrates what the machine thinks are important textual features are for identifying impact investment from an article.

4) Accuracy Testing: We divided our dataset into 2 partitions, a training partition, that consisted of 80% of the dataset, and a testing partition, that consisted of the remaining 20%. After the training error statistic had a reached a minimum, we found that the model that was built was 92% accurate. So, it is safe to say that the model does not suffer from an overfitting problem.

To use this model with an article outside our dataset, the preprocessing procedure mentioned in step 2 is needed for accurate predictions.

IV. Conclusion

This paper demonstrates the usage of state-of-the-art technology such as Natural Language Processing and Artificial Intelligence to make financial news more relevant and enable easy decision making. Future work will include enhancing the web scraping process using the keywords found in the initial article analysis.

Our method currently relies on extractive summarization which provides relevant summaries based on what words are repeated. This may occasionally lead to loss of relevant information. One approach to solving this problem is to use abstractive summarization [9]. This will enable our program to provide relevant information in a summary that mimics a summary written by a human.

In our work, sentiment analysis is currently limited to English articles. In the future, we will expand sentiment analysis to other Languages by training Word Nets for languages other than English. Sponsored news may affect the analysis of sentiment. We will also expand the sentiment analysis to identify sponsored news and to adjust the sentiment of the article accordingly.

Acknowledgment

The authors would like to thank employees of NeuralTechSoft for their helpful criticisms on earlier versions of the paper.

References

[1] E. D. Liddy, "Natural language processing," 2001.
[2] S. Russell, P. Norvig, and A. Intelligence, "A modern approach," *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, vol. 25, no. 27, pp. 79–80, 1995.

[3] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The azfin text system," *ACM Transactions on Information Systems (TOIS)*, vol. 27, no. 2, p. 12, 2009.

[4] A. Mahajan, L. Dey, and S. M. Haque, "Mining financial news for major events and their impacts on the market," in *Web Intelligence and Intelligent Agent Technology, 2008. WI-IAT'08. IEEE/WIC/ACM International Conference on*, vol. 1. IEEE, 2008, pp. 423–426.

[5] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[6] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[7] J. Ren, S. D. Lee, X. Chen, B. Kao, R. Cheng, and D. Cheung, "Naive bayes classification of uncertain data," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 944–949.

[8] R. Mihalcea, "Language independent extractive summarization," in *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*. Association for Computational Linguistics, 2005, pp. 49–52.

[9] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv preprint arXiv:1705.04304*, 2017.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[11] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *CoRR*, vol. abs/1602.03606, 2016. [Online]. Available: http://arxiv.org/abs/1602.03606

[12] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining." in *Lrec*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[13] E. Keogh, "Naive bayes classifier," *Accessed: Nov*, vol. 5, p. 2017, 2006.

[14] M. Thelwall and P. Sud, "Webometric research with the bing search api 2.0," *Journal of Informetrics*, vol. 6, no. 1, pp. 44–52, 2012.

[15] S. Impact and I. TASKFORCE, "Impact investment: The invisible heart of markets," 2014.

[16] W. J. Wilbur and K. Sirotkin, "The automatic identification of stop words," *Journal of information science*, vol. 18, no. 1, pp. 45–55, 1992.