Sentiment Analysis in Financial Domain

Shreyas. K. V\textsuperscript{1}, Sudarshan R. Kulkarni\textsuperscript{2}, Sumanth. S. A\textsuperscript{3}, Suraj Vashista. B. K\textsuperscript{4}, Shashank. S\textsuperscript{5}
Department of Computer Science & Engineering
National Institute of Engineering, India

Abstract:
The role of sentiment analysis in the financial applications in recent years cannot be understated. The sentimental aspects in news texts will undoubtedly result in changes in the prices, volume of trades, and major risks of financial subjects. The existing research in this field mainly focus on identifying polarity (e.g. positive, neutral or negative) using conventional methods and not giving much attention to the certainties regarding the financial subject. However, recent developments in deep learning allows us to experiment with CNNs which were previously thought to be unsuitable for this domain. Further, we can experiment with RNNs and their various forms. In this paper, we present a suitable neural network architecture for sentiment analysis in the financial domain wherein we experiment on real-world datasets.

Keywords: Sentiment analysis, Financial Documents, Polarity.

I. INTRODUCTION

Financial documents or news provides insight about the financial status of a business. Appropriate financial data is depicted in an organized fashion such that it is simple to decipher. These may consist of income statements, profit and loss statements, balance sheets, etc. They are the industry standard method of conveying financial status of the company/organization to external parties.

Financial documents are useful due to the reasons listed as follows:
1. To evaluate the ability of a business enterprise to generate revenue and plan their uses accordingly.
2. To determine whether a business has the capability to pay back its debts and have a faster turn-over.
3. To gain valuable insight based on the various business transactions that are disclosed by the organization along with the statement.

An efficient sentiment analyser model is used to accurately gauge the sentiment of the document in question. This consists of feature extraction and vectorization through tf-idf and Word2Vec techniques, after which a classifier is trained on the obtained feature vectors. These features are analysed to find the discriminatory behaviour between the various sentiments using natural language processing. The end result is to develop a product that produces reliable and accurate sentiment of the given document without any bias whatsoever. This can be achieved by exploring various natural language processing techniques to reduce any inherent biases in the said document.

II. RELATED WORK

The essence of sentiment analysis is the categorization of sentiment polarity. Given a piece of text document, a sentiment polarity is determined based on the categories specified (positive, neutral or negative). Some of the major findings in this domain are briefly discussed in this section of the paper.

[19] proposed classifiers such as SVM, Naive Bayes, Maximum Entropy and it was determined that when they were evaluated on unigram and bigrams, the accuracy of all the classifiers were around 80%. While most of the methods use hand-crafted features, some approaches rely on lexicons with words and their polarity score. These approaches map the words polarity score and compute the sentiment of the document. While these methods work relatively well for other domains, it can't be used for the financial domain where the vocabulary is very esoteric in nature, making hand crafted feature unviable.

[20] proposed a Deep CNN architecture after stating that traditional NLP means of hand-crafted features are inferior as they are over specific as well as time consuming compared to deep learning methods. CNNs in particular have the ability to extract features which are general in nature and flexible, thus they can be used for almost any domain. Thus, it has been found that ConvNets inherently take into account the ordering of words, and with the help of word vectors, can easily represent a word in a spatial context.

[21] proposed an attention driven architecture consisting of Bidirectional GRUs and Word embeddings for the word, sentence, and document levels. Further, queries can be inserted along with the document to provide a more contextual prediction.

[22] proposed a model using financial and non-financial performance indicators to improve the existing systems in the domain. They present a hierarchical sentiment classifier model based on association rule mining to predict the sentiment.

III. IMPLEMENTATION

An overview of sequential steps and techniques commonly used in sentiment classification approaches is shown in the figure.
below. In this model, information is used as part of a feature set which leads to sentiment classification on a dataset.

**Tokenization**
Tokenization refers to the process wherein a body of text must be broken down into its constituent tokens and the punctuation present in the text is removed. A token may be defined as the smallest constituent unit of a sequence of characters. For instance, a word could be taken as a token for a sentence, a sentence could be considered a token of a paragraph, and so on.

**Pre-processing Text and Stop-word Removal**
Pre-processing of data is the process of preparing and cleaning the data of the dataset for classification. Reducing the noise in the text leads to performance increase in the classifier. A stop word is a word that carries no significance in the sentiment polarity of a sentence, thus they can be removed. A stop word list contains such words. Some common stop words in the English language are ['a', 'the', 'of', 'I', 'it'] and so on.

**Embedding and Vectorization**
The training data is split into its constituent tokens after which each word must be mapped to a certain value or a vector. TensorFlow has a convenient way to handle tokenization using the Tokenizer function. The vectors for each word can be formed using tfidf, bag of words or word-to-vec. Additionally, a pre-trained embedding index could also be used such as Glove. The embedding dimensions can either be 50, 100, 200 or 300.

**The different techniques explored include:**
1. Bag of Words – creating a basic frequency list of various tokens in all of the documents.
2. Tf-idf – creating a weight for each lexicon using a variance formula.
3. Word2Vec – creating a vectored representation of each word in a certain number of dimensions.

**Classifiers**
The classifiers used in this research include –
1. Support Vector Machine (SVM)
2. Decision Tree
The accuracy of various models as trained on Amazon product review dataset is as shown in the following figure.

V. CONCLUSION

The paper proposed experiments on various combinations of vectorization techniques and classifiers. We can infer that given a small dataset of around 3000 examples, non deep learning techniques such as SVM, decision trees and naïve Bayes performed well with the accuracies as shown in the table. However, deep learning models such as CNN, LSTM and GRU performed subpar with coin-toss accuracy. Experimentations were done on amazon product reviews dataset from Kaggle which had 3.6 million examples for training these models then performed excellently with accuracy above 94%. Thus, we can conclude that had the deep learning models been supplied enough data, a far better accuracy could have been obtained for the financial domain dataset.

VI. ACKNOWLEDGMENT

This research was supported by National Institute of Engineering. We thank our guides Shashank S, Suhas S and Abhinandan S P who provided insight and expertise that greatly assisted the research of this paper.

VII. REFERENCES

[2]. Author links open overlay panelRitu Yadava1A. Vinay Kumar AshwaniKumar, News-based supervised sentiment analysis for prediction of futures buying behaviour.
[4]. Ashok Tholusuri, Manish Anumala, Bhagyaraj Malapolu, G. Jaya Lakshmi, Sentiment Analysis using LSTM.
[6]. Zohreh Madhoushi, Abdul Razak Hamdan, Suhaila Zainudin, Sentiment analysis techniques in recent works.
[7]. https://towardsdatascience.com/another-twitter-sentiment-analysis-bb5b01ebad90
[9]. https://appliedmachinelearning.blog/2017/02/12/sentiment-analysis-using-tfidf-weighting-pythonscikit-learn/
[10]. https://www.kaggle.com/divsinha/sentiment-analysis-countvectorizer-tfidf
[12]. Xi Ouyang, Pan Zhou, Cheng Hua Li, Lijun Liu, Sentiment Analysis Using Convolutional Neural Network.
[14]. https://www.kaggle.com/bittlingmayer/amazonreviews
[15]. Ling Luo1, Xiang Ao1, Feiyang Panl, Jin Wang, Tong Zhao, Ningzi Yu and Qing He, Beyond Polarity: Interpretable Financial Sentiment Analysis with Hierarchical Query-driven Attention.


[20]. Dario Stojanovski, Gjorgji Strezoski, Gjorgji Madjarov, and Ivica Dimitrovski. Twitter sentiment analysis using deep convolutional neural network

[21]. Ling Luo1, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu and Qing He. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention

[22]. Srikumar Krishnamoorthy. Sentiment analysis of financial news articles using performance indicators