Sentiment Analysis And Topic Modeling of Stock Market Data

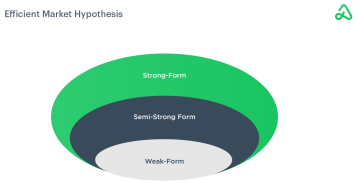
Akhil Bandhu Anand
Computer Science Department
Southern Utah University
Cedar City, United States
akhilanand@suumail.net

***Abstract*—This paper covers an initial increment into stock market sentiment analytics. The purpose of this project is to explore the how good the models can get with predicting sentiment, the project will be an initial platform into modeling sentiment trading. If the models defined in the paper can get good enough in predicting sentiment, technical, fundamental, and algorithmic trading can be modelled with sentiment scores from this paper as a variable. Previous studies have shown that investor sentiment indicators can predict stock market change. This paper will be a mixture of previous work done in this field and then some more.**

I. Introduction

Sentiment, that is one factor that may give financial modeling an edge on the market at least that is what the theory for this project is. To gain an understanding of financial markets we must first look at the Efficient Market Hypothesis (EMH). One fallacy of this theory is that all investors act rationally and as one can gather from intuition of the market, people don't always act rationally especially when finances are involved. Coming back to the theory, there are three factors in the EMH, the first one being weak form hypothesis. If one believes in this form of the hypothesis, the investor can never gain an edge on the market through technical analysis, that is, looking for patterns and trends in historical data to predict future stock market movements. The second leg of the theory is the semi-strong form of the hypothesis, investors cannot get an edge on the market through fundamental analysis, that is, by examining related economic and financial factors such as the balance sheet, strategic initiatives, micro-economic indicators, and consumer behavior, fundamental analysis attempts to identify a security's intrinsic value. And the final leg of the theory, the strong form of the hypothesis. This part of the theory highlights that investors cannot gain an edge on the market through insider trading, gaining information on a security before said information is made public. All this is possible because the theory states that a security's price reflects all the above highlighted information as soon as it is available to anyone in the market. Basically there is no way of gaining an edge on the market!

This paper will explore the analysis of sentiment, it won't reach the level of analysis needed for the bigger project in hand. This analysis is going to be a stepping stone for a polarity score to be used in the making of a financial trading



model. The goal of this analysis is to be able to at least get a good hand at predicting the sentiment of the market using supervised and unsupervised learning methods.

The use of sentiment analysis (SA) is increasingly used to determine the feelings of social media users towards a subject. The most popular way of performing sentiment analysis is data mining. Our aim is to use Deep Learning and Supervised Learning technologies to assess investors' emotions. Our central idea is to adopt Deep Learning to determine investors' expectations about the price of stocks and the overall market based on their messages. The reason why we select Deep Learning methodology rather than data mining is that in data mining, identifying features and selecting the best of those features is the most challenging task to undertake especially in a Big Data.

In contrast to data mining, a Deep Learning model, learns features during the process of learning. Deep Learning algorithms lead to abstract representation, as a result, they can be invariant to the local change in the input data. In addition, Big Data problems including semantic indexing, data tagging, and fast information retrieval can be addressed better with the aid of Deep Learning. Deep Learning provides the opportunity to use a simpler model to accomplish complicated Artificial Intelligence tasks.

The remainder of this paper is organized as follows: "Related Work" section we look at preivious work done in the field of sentiment analysis and the methods employed therein; "Methodology", in this section of the paper, we will look at the methodologies used in this paper. This paper has an emphasizes on the bag of words method and sentiment predictions but there is a section of topic modeling, especially

Latent Dirichlet Allocation, Word Clustering, Text mining, and Polar Subjectivity. In this section, we will also explore the data set and how it is prepared for the rest of the analysis. And the final part of the paper will discuss the results and conclusions from this project.

II. Related Work

Following is some of the work that has been done in the area.

Pandey et al. (2017) suggested a new "metaheuristic scheme" based on the K-mean algorithm and the Cuckoo search method. The emotive information from the "Twitter dataset" was used to find the appropriate CH (Cluster heads). The suggested model's accuracy was tested on a variety of "Twitter datatsets" and compared to other optimization approaches such as PSO ("Particle Swarm Optimization"), DE ("differential evolution"), CS ("cuckoo search"), and two n-gram schemes [1].

Wehrmann et al. (2017) presented a "language-agnostic translation-free" approach for "Twitter Sentiment analysis" using CNN. The polarity of tweets have been examined which has been written in various languages. It has been analyzed that the structure of deep neural network required less learner constraints. The experiment has been carried out on four different languages such as English, german, Portuhuese and Spanish [2].

Sosa and P. M. (2017) proposed an integration of two NN (Neural networks) such as CNN (Convolution "Neural Network") -LSTM ('Long short term memory') and LSTM-CNN to perform sentiment analysis on Twitter data. For training 10,000 tweets have been tested that consists of positive and negative tweets equally. The average accuracy obtained by using LSTM-CNN is high about 75.2% [3].

To discover the feeling level of the News data, Shahare et al. (2017) offered techniques such as 'Naive Bayes' and 'Levenshtein algorithm.' The author uses different levels of processing to determine the emotion words. The author expertly extracts emotion from either a tiny or a large text. The author can also tell the difference between news events and text data [4].

This paper's experiments are focused on market sentiment. Market mood, according to [5], is the overall prevalent sentiment of investors in terms of anticipating price development in a market. This attitude is influenced by a variety of elements, including global events, history, economic news, seasonal conditions, and so on. Sentiment analysis, often known as opinion mining, is the practice of extracting a writer's attitude from source materials using natural language processing methods.

Multiple research initiatives have used supervised classification methods such as Support Vector Machines [7], Nave Bayes [6], or ensembles [7, 8] to perform sentiment analysis. The bag-of-words [52] paradigm is commonly used in machine learning approaches. A text is represented as a collection of its words in the bag-of-words model, irrespective of the sequence of those words in their sentences.

Loughran and McDonald's [10] work is one of the most well-known in this subject. They created a financial lexicon and manually created six-word lists using the US Securities and Exchange Commission portal from 1994 to 2008, including positive, negative, litigious, uncertainty, model strong and model weak.

III. Methodology

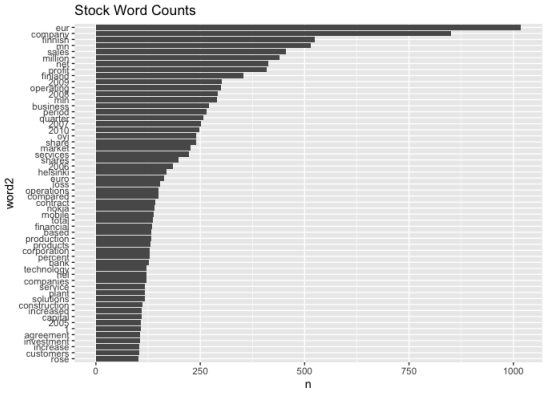
A. Dataset

For this paper, The first resource that would have provided a comprehensive dataset would have been StockTwits Inc. StockTwits is a financial social media platform that was founded in 2009. StockTwits users can get information on the stock market, such as current stock prices, price movement, stock exchange history, buying or selling recommendations, and so on. Furthermore, as a social network, it allows stock market traders to share their experience. Unfortunately, I was not able to gain access to their platform to gather a dataset. I was lucky enough to find a few datasets on Kaggle. Kaggle is a Google subsidiary, an online community of data scientist and machine learning practitioners, this community shares code and datasets about various different problems around the world.

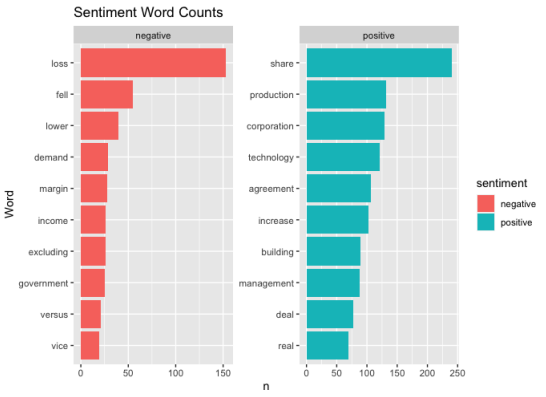
In this experiment, I was able to find two different data sets. This first called "stock data.csv", this data set had two columns and 5,791 rows. The first column was a Sentiment score, "positive" and "negative", the other column was the actual text, this dataset was gathered from multiple twitter handles talking about market sentiment and economic news (web scraping). The next data set called "all data.csv", also had two columns one of Sentiment and one of the actual text. This data set had three categories "positive", "negative", and "neutral". Both these data set were combined to get a final dataset over 10,000 rows with three different categories of sentiment.

B. Text Mining

The first part of the analysis of this dataset is more of an exploratory data analysis, that is cleaning up the data and making some graphs to see how the data is structured. This next graph and word cloud is what the dataset looks like without any cleaning, without removing any kind of stop words or words like "eur" which will not be helpful in an american market.



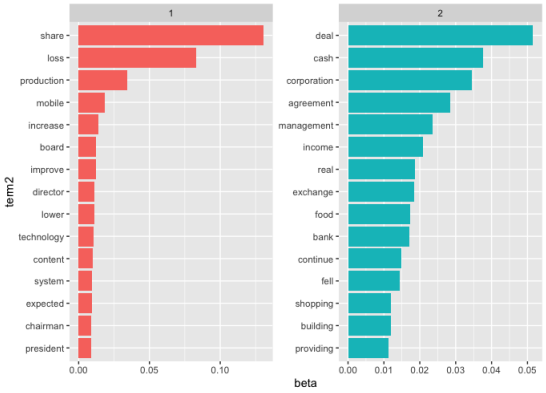
Now, we will do some cleaning on the dataset and create some better indicators of the words look like. To do this, we used the NRC lexicon. The NRC Emotion Lexicon is a list of English words and their associations with eight basic emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were manually done by crowdsourcing. After doing the cleaning and joining with the NRC lexicon, this next graph shows the top 10 positive and negative words in the dataset. The top negative word is ”loss” and the top positive word is ”share” which makes sense intuitively.



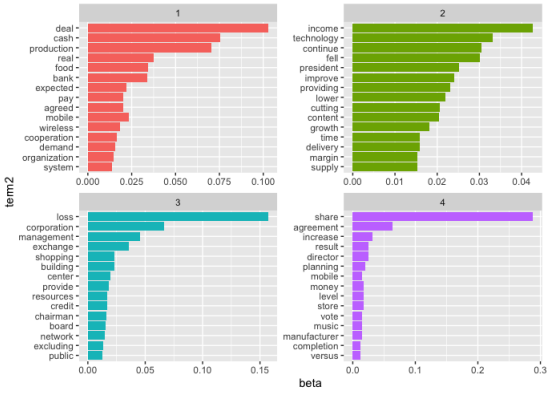
The next thing we will cover is some unsupervised learning, that is, some topic modeling. The method that this paper will cover is called Latent Dirichlet Allocation (LDA).

C. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a method that uses a corpus of text to represent a document as a random mix of hidden themes [11]. In natural language processing, the latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. This next plot is from an LDA model that outputs and creates general topics through word probabilities.



This next plot is creating 4 topics through LDA modeling. These topics are very subjective and the reader's intuition will lead to the creation of the topic.

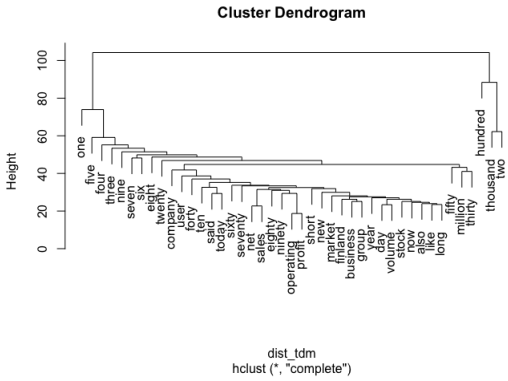


D. Word Clustering

Word clustering is a technique for dividing a set of words into subsets of semantically comparable terms. It is rapidly being utilized in a variety of NLP applications, from word sense and structural disambiguation to information retrieval and filtering.

A dendrogram is a visual representation of the hierarchical relationship between items. It's most often produced as a result of hierarchical clustering. A dendrogram's main purpose is to figure out the best approach to assign objects to clusters.

Hierarchical cluster analysis (also known as hierarchical clustering) is a type of cluster analysis in which the goal is to group objects or records that are "near" to one another. This next plot is a hierarchical cluster plot for the dataset, this shows us the word association for all the words.



E. Supervised Learning

Now, let's pivot into actually doing some classification and modeling

1) Data Examples: Here are some examples of the text data, this will explain why further cleaning needs to be performed to get the data ready for some predictive modeling.

- "clothing retail chain seppl s sales increased eight one thousand five hundred fifty two operating profit rose three hundred eleven one hundred seventy one two thousand four"
- "my 526.20 bid not hit in aapl so i am expecting at least a move to 540 today :)"
- "et the angry, anecdotal comments begin: Why I still think BlackBerry IM are doomed. IMM"
- "HB.A - ong 73.67. Trailing Stop 71.92 from 3 prior Stops of 70.86, 68.64 and 62.41 -"
- "VOC from watch list added Tuesday has now run over the 50-day EMA vol increased to 86%"

2) Pre-Processing: Some of the messages had a very weird symbols in them refer to the examples given above. Some preprocessing steps are performed to clean the text in the dataset, this was done after the above mentioned processing to get the data ready for modeling:

- Text with symbols were removed.
- Stopwords were removed, Examples: any random english words, "finnish", "eur", "will", "aap", "mn", "oyj".
- All the the words were converted to lowercase, this is because models can be very sensitive to capital letters.
- All URLs were removed. Most URLs are short URLs and located at the end of the text.
- All punctuation was removed.
- All whitespace was removed from the text.
- The abbreviations were made into two separate words, like "hadn't" to "had not"
- Symbols were replaced from the text.
- All numbers were replaced with actual text. Example: "1" to "one", "5000" to "five thousand".

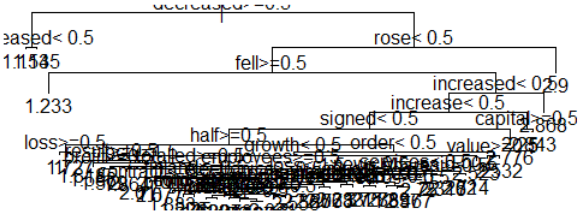
"Corpus" is a collection of text documents. VCorpus in tm refers to "Volatile" corpus which means that the corpus is stored in memory and would be destroyed when the R object containing it is destroyed. After all the cleaning mentioned above was completed, here is a word cloud that represents the most used words in the dataset.

3) Logistic Regression: Logistic regression is a statistical model that uses a logistic function to represent a binary dependent variable in its most basic form, though there are many more advanced variants. Logistic regression (or logit regression) is a technique for estimating the parameters of a logistic model in regression analysis (a form of binary regression). At first, we examine the results from a simple logistic regression at predicting sentiment. For a baseline accuracy, we will take the dominant category. In our testing dataset, out of the three categories, the dominany category is positive sentiment. There are 757 1's in the data making the baseline accuracy 47.49%. This next table is from a logistic

```
Confusion Matrix and Statistics

          Reference
Prediction  -1   0   1
        -1  62   4  26
         0  76  20  70
         1 268 407 661

Overall Statistics

               Accuracy : 0.4661
                 95% CI : (0.4414, 0.491)
    No Information Rate : 0.4749
    P-Value [Acc > NIR] : 0.7664

                  Kappa : 0.0451

 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: -1 Class: 0 Class: 1
Sensitivity            0.15271  0.04640   0.8732
Specificity            0.97475  0.87446   0.1935
Pos Pred Value         0.67391  0.12048   0.4948
Neg Pred Value         0.77097  0.71218   0.6279
Prevalence             0.25471  0.27039   0.4749
Detection Rate         0.03890  0.01255   0.4147
Detection Prevalence   0.05772  0.10414   0.8381
Balanced Accuracy      0.56373  0.46043   0.5334
```

regression model applied onto the dataset. As we can see the accuracy from the model is 46.61%, which is not above the baseline, but the rebalanced accuracy from the class statistics is 51.92%.

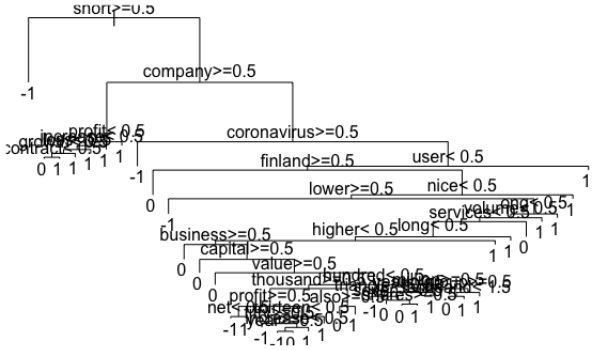
4) ANOVA: Analysis of variance (ANOVA) is a collection of statistical models and associated estimate processes (such as "variation" among and between groups) that are used to examine variations in means. Ronald Fisher, a statistician, invented ANOVA. ANOVA is based on the law of total variance, which divides observed variance in a variable into components attributed to various causes of variation. Our model shows that ANOVA does much better than the logistic regression and also the baseline for this dataset. The plot shows how the ANOVA model makes it's decisions throughout the dataset. This ANOVA model does make a decision tree plot since we us CART modeling which is discussed in the next point. The accuracy of the ANOVA model was 69.79%



anova_class			
	-1	0	1
-1	28	53	9
0	5	397	29
1	8	115	81

5) CART Modeling: A Classification And Regression Tree (CART) is a predictive model that shows how the values of an outcome variable can be predicted based on the values of other variables. A CART output is a decision tree in which each fork represents a split in a predictor variable and each end node represents an outcome variable prediction. We try to anticipate a class label with classification. In other words, classification is used to solve situations where the output (target variable) has a limited range of values, such as whether or not it will rain tomorrow. While this is going on, regression is being utilized to predict a number label. This means that your output, such as a house price, can take an endless number of values. The next plot and table show the statistics from the CART model applied to the dataset. The accuracy of the CART model is slightly improved over the ANOVA model with a prediction accuracy of 70.62%

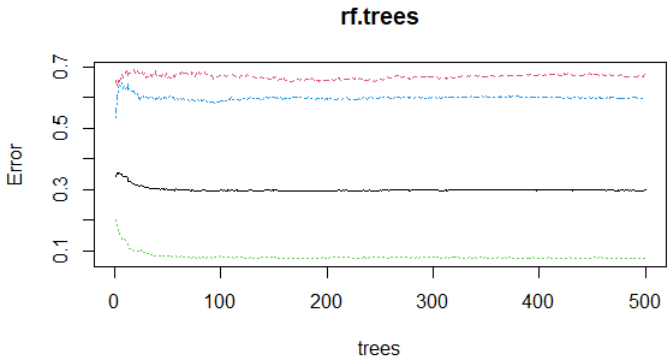
predictclass			
	-1	0	1
-1	28	42	20
0	7	413	11
1	10	123	71



6) Random Forest: Random forests, also known as random choice forests, are an ensemble learning method for classification, regression, and other tasks that works by building a large number of decision trees during training. For classification tasks, the random forest's output is the class chosen by the majority of trees. The mean or average prediction of the individual trees is returned for regression tasks. For our modeling purposes, we ran two tests, one predicting the sentiment on the training dataset to see how the model is doing and the other, predicting the sentiment of the test dataset. The prediction accuracy on the training dataset was 82.44% which came down to 71.31% accuracy on the testing dataset which is much improved from the previous models. These next two tables show the training and testing confusion matrices respectively. As you can see that the prediction on the training data are much higher then the test. There is not a worry for too much overfitting because of the size of the matrix the models are predicting on. The plot shows how the error rate of the random forests reduced over time.

rf_train_pred			
	-1	0	1
-1	292	183	39
0	15	2404	29
1	15	443	702

rf_preds			
	-1	0	1
-1	29	41	20
0	7	403	21
1	12	107	85



7) Naive Bayes: In statistics, Naive Bayes classifiers are a group of straightforward "probabilistic classifiers" based on Bayes' theorem and strong (naive) independence assumptions between features (see Bayes classifier). They are one of the most basic Bayesian network models, but when combined with kernel density estimation, they can achieve great levels of accuracy. Unfortunately, for this dataset, the Naive Bayes classifier did not do that well only being able to achieve and accuracy of 47.03%. If we recall, the baseline accuracy for this dataset was 47.49%.

8) Support Vector Machines: Support-vector machines (SVMs, also known as support-vector networks) are supervised learning models that examine data for classification and regression analysis in machine learning. For the purpose of this analysis, Support Vector Machines actually did quite well. the next table shows the prediction output from an svm classifier on the testing dataset. The model had an accuracy of 78.89% on the training dataset which reduced to 70.76% for the testing dataset which is not bad at all.

	svm_pred		
	-1	0	1
-1	25	48	17
0	3	408	20
1	5	119	80

IV. Results And Conclusion

Finally, let's discuss some of the results from the reseatch done in this paper. First, let's talk about the initial text mining and topic modeling section of this paper. To get the top 10 positive and negative words in the dataset using the NRC lexicon was interesting because intuitively you would think that those are the top 10 in each category. It was very interesting to do some Latent Dirichlet Analysis, create some topics that kind of model the entire dataset. From the topics we created, it can be seen that there are 4 very distinct categories in the dataset that people talk about. Production, Technology, Management, and the Manufacturing sectors seem to be the most talked about in the text. Now, the results from word clustering, it can be seen that numbers are very closely related to each other and it gives a good idea as to how the sequence of words may affect the sentiment of a tweet. And now finally, let's talk about some supervised learning results. The logistic regression did not do that well with an accuracy of 46.61% now even above the baseline accuracy, this is understandable since logistic models are usually used to model binary problems but this dataset has three categories that need to be predicted. The next best model, a surprising result here was the Naive Bayes Classifier model with an accuracy of a meager 47.03%, still not better than the baseline. The next best model was the ANOVA model with an accuracy of 69.79%, 22% better than the baseline! This can be foreseen since ANOVA is a way of extracting features from the dataset and hence should do better than a logistic regression at least. A slightly better model was the CART model, a decision tree based model which was quite surprising. This model had an accuracy of 70.62%. Support Vector Machines did a tiny bit better with an accuracy of 70.76%. And finally the behemoth of a model with a random forest, being a tree based model but running a lot of different trees, it is understandable that this would do much better than the CART models at least, these models had an accuracy of 71.31% on the testing dataset. 24% better than the baseline accuracy. For the reseach conducted in this paper, this is a very good result and first step into the financial stock market sentiment analysis.

References

[1] A. Chandra Pandey, D. Singh Rajpoot and M. Saraswat, "Twitter Sentiment Analysis using Hybrid Cuckoo Search Method", Information Process Managemen, vol. 53, no. 4, pp. 764-779, 2017.
[2] J. Wehrmann, W. Becker, H. E. L. Cagnini and R. C. Barros, "A Character-Based Convolutional Neural Network for Language-Agnostic Twitter Sentiment Analysis", proceedings International Conference on Neural Networks, pp. 2384-2391, 2017.
[3] P. M. Sosa, "Twitter Sentiment Analysis using Combined LSTM-CNN Models", pp. 1-9, 2017, [online]
[4] F. F. Shahare, "Sentiment Analysis for the News Data Based on the Social Media", Proceedings 2017 IEEE International Conference on Intelligent Computer Control Systems, pp. 1365-1370, 2017
[5] Market Sentiment. http://www.investopedia.com/.
[6] Saif H, He Y, Alani H. Semantic sentiment analysis of Twitter. The semantic Web-ISWC 2012. Berlin: Springer; 2012. p. 508–24.
[7] Steinwart I, Christmann A. Support vector machine. Berlin: Springer; 2008.
[8] Silva N, Hruschka E, Hruschka E. Tweet sentiment analysis with classifier ensembles. Decis Support Syst. 2014;66:170–9.
[9] Fersini E, Messina E, Pozzi FA. Automatic construction of financial semantic orientation lexicon from large scale Chinese news corpus. Decis Support Syst. 2014;68:26–38.
[10] Loughran T, McDonald B. When is a liability not a liability? Textual analysis, dictionaries. J Finance. 2011;66:35–65.
[11] S. J. Putra, M. A. Aziz, and M. N. Gunawan, "Topic analysis of Indonesian comment text using the latent Dirichlet allocation," 2021 9th International Conference on Cyber and IT Service Management (CITSM), 2021.