iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** X **Month of publication:** October 2021

DOI: https://doi.org/10.22214/ijraset.2021.38345

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue X Oct 2021- Available at www.ijraset.com

Summarization and Sentiment Analysis for Financial News

Anusha Kalbande
Student, Cummins College of Engineering for Women, Pune

Abstract: ***Data is growing at an unimaginable speed around us, but what part of it is really useful information? Business leaders, financial analysts, stock market enthusiasts, researchers etc. often need to go through a plethora of news articles and data every day, and this time spent may not even result in any fruitful insights. Considering such a huge volume of data, there is difficulty in gaining precise, relevant information and interpreting the overall sentiment portrayed by the article. The proposed method helps in conceptualizing a tool that takes financial news from selected and trusted online sources as an input and gives a summary of the same along with a basic positive, negative or neutral sentiment. Here it is assumed that the tool user is familiar with the company's profile. Based on the input (company name/symbol) given by the user, the corresponding news articles will be fetched using web scraping. All these articles will then be summarized to gain succinct and to the point information. An overall sentiment about the company will be portrayed based on the different important features in the article about the company.***

Keywords: Financial News; Summarization; Sentiment Analysis.

I. INTRODUCTION

In the financial domain, even the smallest piece of information is important. A simple business announcement can either increase the stock price 2 times or even be the reason for its crash. Business leaders, financial analysts, stock market enthusiasts, researchers etc. often need to go through a plethora of news articles and data every day for their predictions. But this time spent may not result in any fruitful insights often. Considering such a huge volume of data, there is difficulty in gaining precise, relevant information and even interpreting the overall sentiment portrayed by the article. Hence arises the need for some technology that can preprocess all this raw data to give out useful information.

The proposed method helps in conceptualizing a tool that takes financial news articles from selected and reliable online sources as an input and gives the user:

A. A Succinct Summary Of The News Article

The summary would consist of only relevant and important subject matter from the article. Irrespective of the length of the input article, the summary would be approximately 5-7 lines each time. In many cases, this summary can be an equivalent of the direct facts in the article along with its actual title.

B. An Overall Basic Sentiment Portrayed By The Article

By analyzing all the keywords and phrases in the article, the user can get to know if the article is positive, negative or neutral from a general point of view. While there are many sentiments an article can convey and many complex algorithms to achieve those, this tool focuses on the basic result while leaving the rest of the judgement to the user.

It is expected that the user is familiar with the company's profile for which he/she wants the summarized articles for. Between taking the user's input to a final summary and sentiment, this tool comprises 5 major modules that will help in processing the data at each stage as Figure.1.

II. INPUT AND QUERY PROCESSING

When the user inputs the name of a company, news articles are derived from certain reliable online sources such as The New York Times[1], Forbes[2] and Economic Times[3] by web scraping. The web scraping module is further divided into 2 sub modules - URL fetcher and the site crawler. The URL fetcher will fetch the URLs for the required company or firm from the online source. Since each site has a different HTML structure, a site crawler is necessary to extract only the useful text out of the whole article, by skipping any subtitles or advertisements. The output from this module serves as an input to the next module where this raw data is further processed.

III. DATA PREPROCESSING

Data preprocessing is crucial as it removes unwanted noise. For preprocessing, NLTK (Natural Language Toolkit) and Regular expressions are used throughout. The following preprocessing techniques are applied to the raw news articles:

A. Removal of punctuations, numbers.
B. Tokenization - Sentences are broken down to words/ tokens for better processing. Removal of Stop words - Tokenized words are checked against a built-in list of stop words to remove them. e.g., 'the', 'is', 'are'.
C. Removal of whitespaces.

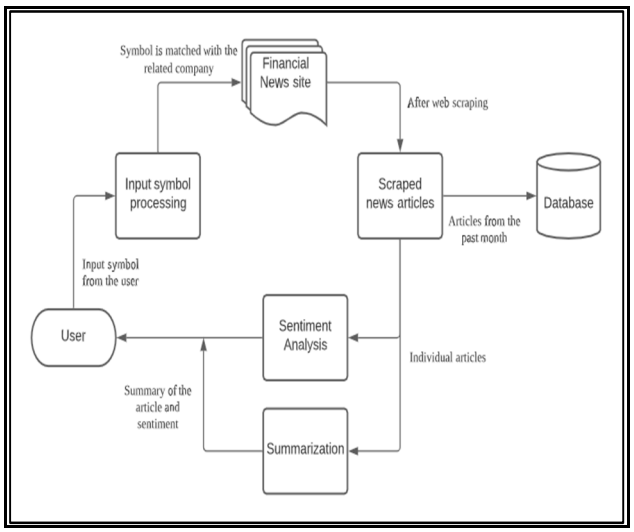


Figure 1

IV. STORAGE

The extracted, preprocessed and summarized data has to be stored somewhere for future use. All the summarized articles with the sentiment values and original URLs are stored in a database, Flask-SQLAlchemy in this case. It is a powerful Object Relational Mapper with the flexibility of SQL. Articles for the past one month are stored for this tool to use; articles older than are deleted by referencing their timestamp.

V. SENTIMENT ANALYSIS

Sentiment analysis consists of:

1) Categorization of opinions and emotions conveyed in an article.
2) Determination of tone- positive/negative/neutral.

The sentiment of the article is calculated prior to the summarization, so that all the keywords and phrases in the article are accounted for. The article's polarity is determined by counting the number of positive and negative words in it. Since the sentiment required is of a simple nature, a single lexicon approach is being used here. For implementation, two csv files, Positive.csv and Negative.csv composed of the most common positive and negative words are created. To ensure that only relevant lexicons from the article are used for sentiment analysis, the *tf*idf* [1] measure is used.

tf = (Number of times term t appears in a document) / (Total number of items in the document)

idf = ln (Total number of documents / Number of documents with term t in it).

Once the important lexicons are identified, they are compared with words from Positive.csv and Negative.csv respectively using *cosine similarity*[2]. An article is classified as neutral if it has the same number of positive and negative neutral words.

VI. SUMMARIZATION

Extractive summarization approach is used in this module, as the user needs only 5-7 sentences from the original article. The Text Rank algorithm in NLP is being used for this step. Text Rank algorithm is analogous to Google's Page Rank algorithm where instead of ranking pages, sentences in a text paragraph are ranked. Following steps are executed in order to get a summary of the article:

A. Similarity Matrix

All the 'n' sentences from the article are arranged into a n*n matrix. The entry simMat[i][j] corresponds to the similarity between the sentences 'i' and 'j' which is calculated using the cosine similarity metric.

B. Text Rank Algorithm

Initially the ranks of all the sentences are assumed to be one. The compression rate[3] is decided for the algorithm. The ranking algorithm starts by picking the first sentence in the matrix. The next sentence's rank will be based on the first's updated rank.
After all the sentences in the article are ranked and arranged in descending order of the rank, the top 5-7 (depending on the compression rate for text-rank algorithm) sentences are grouped together to form the summary of the article.

VII. EVALUATION

Recall-Oriented Understudy for Gisting Evaluation - ROUGE has been used to analyze the performance of the system. ROUGE evaluates the automatic summarization of texts by comparing the system-generated summary with the human generated summaries (Gold - standard summaries). Precision, Recall and F-score are given as a result of the evaluation.
In terms of summaries, a high recall value indicates that most of the words in the reference summary are captured in the system summary. But most of those words might actually be unnecessary in the summary, making it verbose. Recall just by itself cannot be relied upon, so we have precision, which tells how much of the words in the system summary were actually needed. F-measure takes into account both precision and recall and thus is a better way to judge the performance.
The only limitation while using ROUGE for evaluating this project's summary is that the metrics are based on the comparison with human generated summaries which in this case are fellow students and not experts, which could lead to a reduced score.

A. Observations For 65% Compression Rate

1) Average precision = 59.44 %
2) Average recall = 48.5 %
3) Average F-score = 56.74 %

B. Observations For 75% Compression Rate

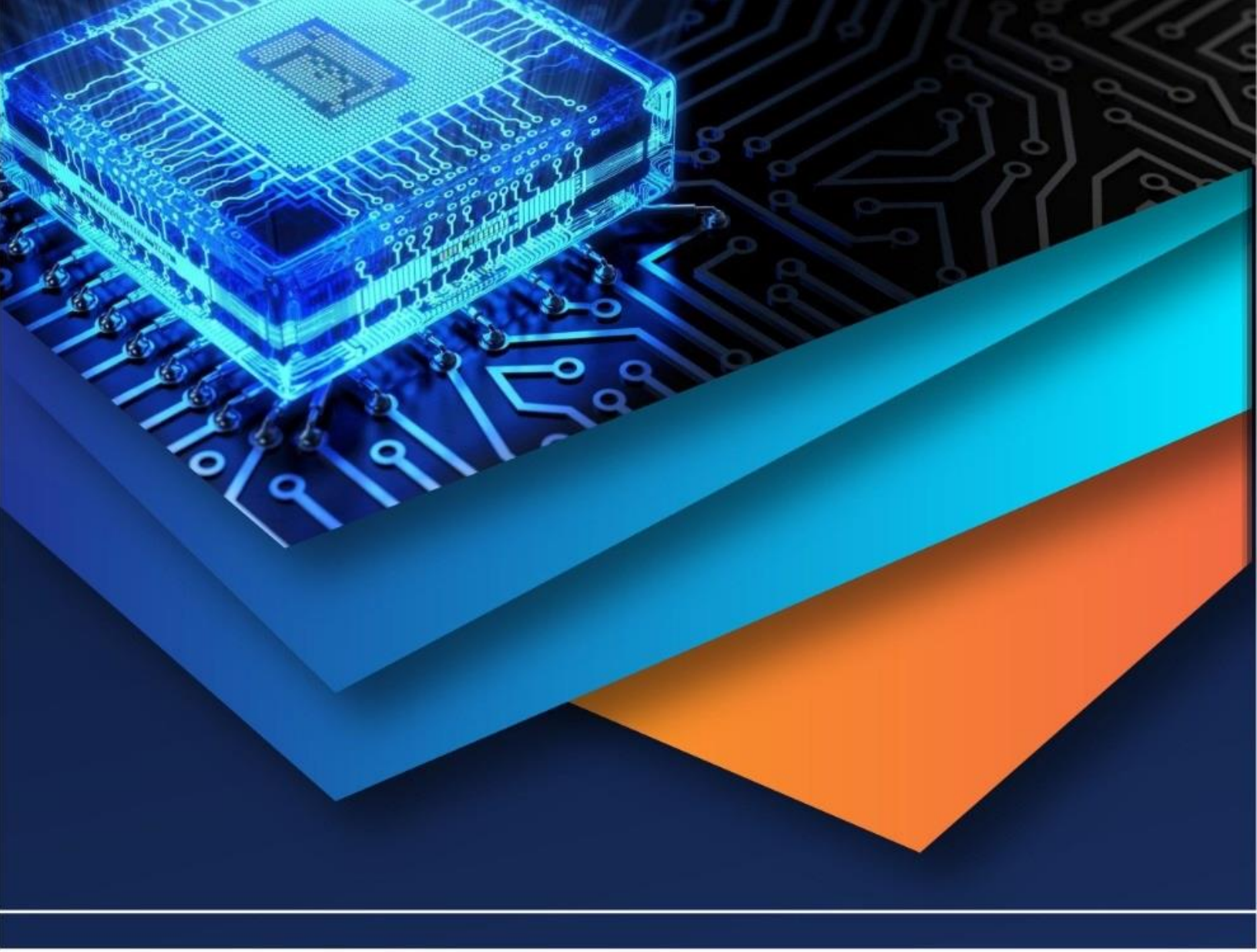
1) Average precision = 57.74 %
2) Average recall = 53.54 %
3) Average F-score = 55.42 %

VIII. CONCLUSION

A. 65% compression rate performs better than 75%.
B. Since more of the original article is retained at a 65% compression rate, the summary is longer than needed sometimes.
C. Thus, if the number of lines has to be less for the summary, the measures would be compromised a bit. There is a tradeoff between performance measures and summary length.

REFERENCES

[1] Katja Filippova, Mihai Surdeanu, Massimiliano Ciaramita, Hugo Zaragoza, "Company-oriented extractive summarization of financial news.", 2009, In COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics:2009 895-903.
[2] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, and Krys Kochut, "Text Summarization Techniques: A Brief Survey.", 2017, In Proceedings of arXiv, USA, 9 pages.
[3] Mirani, Tarun B. and Sreela Sasi, "Two-level text summarization from online news sources with sentiment analysis.", In 2017 International Conference on Networks & Advances in Computational Technologies (NetACT) (2017): 19-24.
[4] P. Krishnaprasad, A. Sooryanarayanan and A. Ramanujan, "Malayalam text summarization: An extractive approach" , 2016, International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, 2016, pp.1-4.
[5] Marco Bonzanini, Miguel Martinez-Alvarez and Thomas Roelleke, "Extractive Summarisation via Sentence Removal: Condensing Relevant Sentences into a Short Summary", 2013, In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, Pages 893-896.
[6] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R, "Sentiment Analysis of Twitter Data.", 2011, In LSM'11 Proceedings of the Workshop on Languages in Social Media Pages 30-38.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY