NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

FACULTY OF SOCIO-ECONOMIC AND COMPUTER SCIENCES

SOFTWARE ENGINEERING

PROJECT PROPOSAL

SITE DEVELOPMENT FOR AUTOMATIC COLLECTION, ANALYSIS AND VISUALIZATION OF COMPANY ETHICAL BEHAVIOR

Roman Solomatin, SE-19-1

Supervisor: PHD, Professor of the Department of Information Technologies in Business Perm HSE, A. V. Buzmakov,

PERM

2023

Abstract

Introduction

Background. The ethics of companies have long been of concern to stakeholders, particularly with regard to their actions in contentious situations and their delivery of customer-centric services. In recent years, there has been a growing emphasis on evaluating the ethical standards of companies, particularly in the banking sector and through the lens of Environmental, Social and Governance (ESG) factors(Murè et al., 2021, Miralles-Quirós, Miralles-Quirós, & Redondo Hernández, 2019, Climent, 2018). The need for such assessments has become increasingly urgent as society continues to grapple with the consequences of corporate misconduct and the broader impact of corporate activities on society as a whole.

Assessing a company's ethical standards is a complex process that involves evaluating various aspects of a company's operations, such as its business practices, policies, and overall culture. In addition to traditional financial metrics, ESG factors play a critical role in determining a company's overall ethical standing. All of this provides valuable insight into a company's ethical standards.

Problem statement. Currently, there are a number of services that claim to evaluate a company's ethics, but these evaluations are often based on court cases and other official records rather than customer feedback. This has resulted in a scenario where individuals must conduct their own research to determine the ethics of a particular company. This research often involves reviewing customer feedback from various websites, which can be time-consuming and may not always provide a comprehensive or accurate picture of a company's ethical practices.

To address this issue, there have been recent calls for the development of a system that would collect and analyze customer feedback from multiple websites to provide a more comprehensive view, and this system could also incorporate other data sources such as financial reports, sustainability reports, and news articles to provide a more holistic view of a company's ethical practices. The system could also incorporate machine learning algorithms, such as sentiment analysis, to analyze customer feedback and extract valuable insights. These insights could then be used to generate a score or rating that provides

an overall assessment of a company's ethical practices. In order to accomplish this task, methods of natural language processing will be used.

Aim and objectives. The aim of this project is to develop a neural network capable of analyzing texts for ethics measurement. To accomplish this goal, the following objectives should be achieved:

1. Analyze existing approaches to ethics measurement
2. Select neural network architecture
3. Collect data to fine-tune and analyze
4. Development of neural network
5. Analyzing texts
6. Ethics calculation

Delimitations of the study. The initial focus of this research will be in collecting data from sites with customer feedback and financial reports of different companies and analyzing them. The analysis will be carried out using neutral language processing algorithms.

Professional significance. The study aims to provide valuable insights into a company's ethical practices through the analysis of customer feedback and other data sources. These insights can help stakeholders make informed decisions about their investments and interactions with companies.

Literature Review

Ethics Measurement

The importance of ethical behavior in business cannot be overstated. As evidenced by various studies (Climent, 2018, Murè et al., 2021), companies that prioritize ethical behavior tend to achieve greater financial success and better business performance over the long term than those that engage in unethical practices.

Evaluating a company's ethical standards can be approached from several perspectives. From the perspective of the company itself, various factors can be considered, including the level of capitalization to ensure that it is not at risk of bankruptcy, the impact it has on the surrounding environment, and the direction of its investments (Har-

vey, 1995). On the other hand, customers may focus on the quality of customer service (Brunk, 2010) and the degree to which a company's services are intrusive (Mitchell, Lewis, & Reinsch, 1992).

It is important to note that evaluating a company's ethics is not a one-time endeavor. Rather, it is an ongoing process that requires continuous monitoring and evaluation of the company's actions, policies and practices over time. This is particularly important given the constantly evolving nature of business conduct and the need to stay abreast of emerging issues and trends. In addition, it is important to consider the broader societal impacts of corporate activities and to evaluate companies not only on their financial performance, but also on their environmental, social and governance (ESG) factors.

Text Analysis

The field of natural language processing has seen a significant advancement in recent years(Devlin et al., 2018, Wang et al., 2018), largely due to the emergence of neural network-based algorithms. These algorithms represent text data in a more nuanced and complex manner, allowing for a deeper understanding of the underlying semantics and meaning. They can help to analyze semantics of texts.

Machine learning algorithms for text analysis have been widely used to extract information from unstructured data using large annotated datasets. Among the various methods used, several algorithms have proven to be particularly effective in this area. These include the bag of words (Harris, 1954), TF-IDF (Jones, 1972), Word2Vec(Mikolov et al., 2013) , ELMO (Peters et al., 2018), GPT (Radford et al., 2019), and BERT (Devlin et al., 2018). Each of these algorithms has unique characteristics that make it well suited for specific applications.

The bag of words model represents text data by assigning a unique number to each word in a document. This method is easy to implement, but does not take into account the order of words in a sentence. On the other hand, the TF-IDF model represents text data by considering both the Term Frequency(TF) in a document and its Inverse Documents Frequency(IDF) in the corpus. This approach can be used to determine the importance

of a word in a given document and is commonly used in information retrieval and natural language processing tasks, but these algorithm do not understand full context of words.

Word2Vec utilizes a vector representation of words, which enables the algorithm to capture the meaning of words in similar contexts. This allows for a more accurate and sophisticated representation of the relationships between words, leading to improved performance in tasks such as text classification and sentiment analysis.

ELMO, GPT, and BERT, on the other hand, are based on the transformer architecture, in which each sentence is represented by a vector of numbers, commonly known as an embedding. This representation allows for a more comprehensive and holistic understanding of the text, as it takes into account the context of the entire sentence or text.

Of these algorithms, BERT is considered to be the most advanced and powerful(Devlin et al., 2018), as it is able to consider the context of the entire sentence or text, whereas GPT and ELMO only consider a one-sided context. This allows BERT to achieve state-of-the-art performance in a wide range of nature language processing (NLP) tasks, including text classification, named entity recognition, and question answering.

For speeding up process of text analysis will be using Sentence-Bert(Reimers & Gurevych, 2019). The superiority of the proposed model over conventional BERT models is due to its innovative approach to sentence embedding comparison. Unlike traditional BERT models, which require recomputation of each pair of sentence embeddings to perform comparisons, this model allows independent comparison of sentence embeddings. This greatly improves computational efficiency, as the following example illustrates.

In traditional BERT models, searching for similar sentences among 10,000 requires 50 million calculations of different sentence pairs, a process that can take up to 50 hours. In contrast, Sentence BERT computes the embedding of each sentence individually before performing a comparison. This results in a significant acceleration of the program execution, reducing the time to only 5 seconds.

Therefore, the proposed model represents a major advance in the field of NLP, enabling more efficient and effective sentence comparisons. This is due to its unique

approach of computing sentence embeddings independently, which provides a distinct advantage over traditional BERT models.

Methods

The purpose of this study is to provide a comprehensive analysis of consumer perceptions of different companies by collecting and analyzing reviews from websites. The first step is to collect reviews using web scraping and application programming interfaces (APIs).

The sentiment analysis of the collected reviews is performed using a fine-tuned Sentence BERT (Reimers & Gurevych, 2019) model. The fine-tuning process is crucial for achieving high accuracy in sentiment analysis, and involves adjusting the parameters of the model to better fit the specific data encountered in this study. The model is trained specifically for the task of analyzing reviews. The fine-tuned Sentence BERT model will classify each review into one of several sentiment categories, such as positive, negative, or neutral, providing valuable insight into the overall sentiment of the reviews.

The final step is to analyze all of the company reviews, resulting in a score for each company based on the reviews. This approach provides a comprehensive assessment of the companies under consideration and a basis for making informed decisions.

Results Anticipated

The primary objective of this study is to collect a dataset of customer reviews for companies and to create and train a model for text analysis. The final outcome of the research will be the analysis of the textual data and the calculation of the ethical standing of the companies under consideration.

Conclusion

References

Brunk, K. H. (2010). Exploring origins of ethical company/brand perceptions—a consumer perspective of corporate ethics. *Journal of Business Research*, *63*(3), 255–262.

Climent, F. (2018). Ethical versus conventional banking: A case study. *Sustainability*, *10*(7), 2152.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162.

Harvey, B. (1995). Ethical banking: The case of the co-operative bank. *Journal of Business Ethics*, *14*(12), 1005–1013.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Miralles-Quirós, M. M., Miralles-Quirós, J. L., & Redondo Hernández, J. (2019). ESG performance and shareholder value creation in the banking industry: International differences [Number: 5 Publisher: Multidisciplinary Digital Publishing Institute]. *Sustainability*, *11*(5), 1404.

Mitchell, W. J., Lewis, P. V., & Reinsch, N. (1992). Bank ethics: An exploratory study of ethical behaviors and perceptions in small, local banks. *Journal of Business Ethics*, *11*(3), 197–205.

Murè, P., Spallone, M., Mango, F., Marzioni, S., & Bittucci, L. (2021). ESG and reputation: The case of sanctioned italian banks. *Corporate Social Responsibility and Environmental Management*, *28*(1), 265–277.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, *abs/1804.07461*.