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

Stakeholders have long been concerned about the ethics of companies, particularly how they behave in contentious situations and how they deliver customer-focused services. The aim of this project is to develop a system for evaluating the ethical standards of companies through the analysis of customer feedback. The system uses fine-tuned Sentence BERT to predict the sentiment of customer feedback and extract valuable insights that can be used to generate a score or rating assessing a company's ethical practices. The main expected outcome of the project is an algorithm capable of accurately analyzing text for ethics measurement. This project can help stakeholders, such as customers, employees, regulators and managers, to assess the ethical practices of companies.

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, Mar, 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 these metrics provide 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 attempts to develop a system that collects and analyzes customer feedback from multiple websites to provide a more comprehensive view, 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 information from them. These insights could then be used to generate a score that provides an overall assessment of a company's ethical practices. In order to accomplish this task, methods of natural language processing (NLP) 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 that is suitable for solving the problem
3. Create program that collects reviews
4. Development of neural network
5. Fine-tune neural network
6. Analyzing texts with neural network
7. 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

It is important to note that evaluating a company's ethics is not a one-time endeavor, but 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 ahead 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(Brunk, 2010), but also on their environmental, social and governance (ESG) factors.

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 to impact ethics, 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 (Harvey, 1995). On the other hand, customers can measure ethics by the quality of customer service (Brunk, 2010) and the degree to which a company's services are intrusive (Mitchell, Lewis, & Reinsch, 1992), these metrics will be considered in this project.

Text Analysis

The field of NLP 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 analyze the semantics of texts in order to assess ethics later.

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).

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 NLP 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 document.

Among these transformer-based algorithms, BERT is considered to be the most advanced and powerful due to its unique approach to text processing (Devlin et al., 2018). Unlike GPT and ELMO, which only consider a one-way context, BERT considers the context of the entire sentence or text, enabling it to achieve state-of-the-art performance in a wide range of NLP tasks, including text classification, named entity recognition, and question answering.

The superior performance of BERT is due to the dual nature of its training techniques. First, it employs a technique known as masked language modeling, in which 15%

of the random tokens (words in sentences) in each sentence are replaced by a special token `[MASK]` and then predicted based on context. In addition, 10% of the tokens are replaced with random tokens, and another 10% are replaced with random words. This approach helps the model understand the relationship between words and their context.

Second, BERT is trained on the next sentence prediction task, which involves predicting whether two sentences will follow each other. To do this, the model is exposed to pairs of sentences, with 50% of the pairs randomly selected from nearby sentences and the other 50% from more distant sentences. This training technique helps the model understand the relationship between sentences, allowing it to make more accurate predictions about the relationships between different sentences in a text, making this algorithm most suitable for text analysis.

To speed up the process of text analysis, Sentence-Bert will be used (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. 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.

Conclusion

In conclusion, the importance of ethical behavior in business has been emphasized by various studies. Evaluating a company's ethical standards is an ongoing process and should consider not only financial performance, but also customer feedback. In the field of NLP, advances in algorithms have led to improved text analysis techniques. Among these techniques, BERT is considered the most advanced and powerful because it considers both word and sentence context in its processing. The use of Sentence-Bert is proposed to speed up the text analysis process.

Methods

The aim of this project is to provide a score based on an in-depth examination of consumer attitudes toward various companies through the systematic collection and analysis of online reviews. This will be achieved through the use of Web scraping techniques, and Application Programming Interfaces (APIs) will be used to achieve this. For this purpose, APIs will be implemented that can collect data and store it in a database, and parsers that would collect data from different sources.

The collected reviews will be analyzed with sentiment analysis methods using a fine-tuned Sentence BERT model (Reimers & Gurevych, 2019). The fine-tuning process is essential to improve the accuracy of the sentiment analysis, and involves adjusting the parameters of the model to better fit the specific dataset used in this study. The model will be trained specifically for the task of sentiment analysis using a methodology similar to the training process of the original BERT model. The model will be fed pairs of sentences and labels indicating whether the sentences belong to the same text.

The fine-tuned Sentence BERT model will classify each review into one of several sentiment classes, such as positive, negative, or neutral, providing a comprehensive understanding of the overall sentiment of the reviews. To do this, the model will be presented with sentences and labels that indicate the sentiment of the sentences as positive, negative, or neutral.

The final stage of the study involves the analysis of all company reviews, resulting in a score for each company based on the aggregated reviews. This approach provides a thorough evaluation of the companies being researched and serves as a basis for making decisions.

Results Anticipated

In this project, customer reviews from different sources will be collected with the use of parsers. For this, parsers will be adapted to each site to extract data, and a program will be responsible for receiving, transforming, and storing the collected information in a database. Then Sentence BERT will be fine-tuned and trained on the collected texts specifically for sentiment analysis. After that, the model will predict scores for each

sentence of the texts, and the scores will be aggregated for each month for each company, reflecting dynamics of changes in customer perception over time.

Conclusion

This project aims to improve the complex issue of assessing a company's ethical standards by developing a neural network to analyze customer feedback. The use of NLP algorithms, such as sentiment analysis, will play a critical role in extracting valuable insights and generating a score that provides an overall assessment of a company's ethical practices. The results of this project can potentially provide valuable information to customers about which companies are best to work with, to managers about the interaction between employees and customers, and to regulators about the state of affairs within companies.

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*.

Mar, M., Miralles-Quirós, J. L., & Redondo Hernández, J. (2019). ESG performance and shareholder value creation in the banking industry: International differences. *Sustainability*, *11*(5), 1404.

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*.

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*.