NATIONAL RESEARCH UNIVERSITY HIGHER SCHOOL OF ECONOMICS

Faculty of Socio-Economic and Computer Sciences

Site Development for automatic collection analysis and visualization of company ethical behavior

PROJECT PROPOSAL

Roman Solomatin PI-19-1

Supervisor

PHD, docent department of Information Technologies in Business HSE Perm A. V. Buzmakov

Perm

2023

TOC

1 Introduction 3

2 Literature Review. 4

2.1 BERT 4

2.2 Sentence BERT 5

2.3 CLIP 6

3 Methods. 7

4 Results Anticipated 7

5 Conclusion. 7

Reference 7

In this paper, an analysis of the ethics of different companies is carried out.

Introduction

The trustworthiness of companies has long been a matter of concern, particularly with respect to their behavior in contentious situations and their delivery of customer-centric services. In recent years, there has been a growing emphasis on assessing the ethicality of companies(Murè, Spallone, Mango, Marzioni & Bittucci, 2021, Miralles-Quirós, Miralles-Quirós & Redondo Hernández, 2019, Climent, 2018), particularly within the banking sector and through the lens of Environmental, Social and Governance (ESG) factors. 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 and the environment.

Currently, there are several services that purport to assess a company's ethics, but these assessments are often based on court cases and other official records rather than on customer feedback. This has led to a situation 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 and accurate assessment of a company's ethical practices. Such a system could be designed to automatically collect and analyze customer feedback from a variety of sources, including social media and review sites. The collected data could then be analyzed using various techniques, such as natural language processing and machine learning, to identify patterns and trends related to a company's ethical practices. The resulting analysis could then be used to develop a more robust and reliable system for assessing the ethicality of companies.

Literature Review

BERT

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that has been shown to be highly effective for natural language processing tasks such as text classification, question answering, and named entity recognition. BERT is based on the transformer architecture introduced in the paper "Attention is All You Need"by Vaswani et al. (2017). The transformer architecture is a neural network architecture that uses self-attention mechanisms to process input sequences in parallel, rather than sequentially as in traditional recurrent neural network architectures.

The BERT model is trained using a technique called "masked language modeling,"in which the model is trained to predict the value of certain randomly masked tokens in a sentence, given the surrounding context. This training process allows the model to learn the relationships between words in a sentence and to represent each word in a high-dimensional vector space called an embedding. These embeddings capture the meaning of the words in a sentence and can be used to represent the sentence as a whole.

BERT can be used to obtain embeddings to assess the tone of reviews about companies. By applying BERT to a dataset of reviews, it is possible to obtain embeddings for each review. These embeddings can then be used to train a classifier to predict the tone of the reviews. The classifier can be trained to predict whether a review is positive, negative, or neutral. Once trained, the classifier can be used to predict the tone of new reviews, providing a reliable and efficient way to assess the tone of reviews about companies.

BERT is not the only algorithm that can be used to obtain embeddings, there are other pre-trained models such as ELMO (Peters et al. 2018), GPT (Radford et al. 2018), Word2vec (Mikolov et al., 2013), TF-iDF (Ramos et al., 2003), and bag of words (Manning et al., 2008) that can be used to obtain embeddings. However, BERT has shown to outperform these algorithms in a variety of natural language processing tasks. For example, BERT has been shown to achieve state-of-the-art performance on a wide range of text classification tasks, question answering, and named entity recognition tasks, outperforming previous models such as ELMO, GPT and Word2Vec.

One of the main reasons for BERT's superior performance is its ability to handle bidirectional context. Unlike ELMO and GPT, which are unidirectional models, BERT is trained to take into account the context both to the left and to the right of a word, resulting in more accurate representations of the meaning of words in a sentence. Additionally, BERT's transformer-based architecture allows for the efficient parallel processing of input sequences, making it faster and more efficient than previous models such as Word2Vec and TF-iDF which are based on sequential architectures.

In conclusion, BERT is a powerful algorithm for natural language processing tasks, it can be used to obtain embeddings that capture the meaning of text, which can then be used to train a classifier to predict the tone of reviews on companies. Compared to previous algorithms such as ELMO, GPT, Word2vec, TF-iDF and bag of words, BERT has shown to be more accurate and efficient. This algorithm provides an effective and efficient way to assess the tone of reviews, allowing organizations to better understand the perceptions of their customers and make data-driven decisions.

Sentence BERT

Sentence-BERT (SBERT) is an extension of BERT specifically designed to generate sentence embeddings. The main difference between BERT and Sentence-BERT is that the latter is trained to predict missing sentences given the surrounding context, whereas BERT is trained to predict missing words in a sentence given the surrounding context.

One of the main advantages of Sentence-BERT is that it can capture the meaning of entire sentences, whereas BERT can only capture the meaning of individual words. This allows Sentence-BERT to better represent the meaning of a text as a whole and to provide more accurate embeddings for text classification tasks, question answering, and named entity recognition.

Another advantage of Sentence-BERT is that it can handle longer input sequences than BERT. BERT is typically trained on sequences of up to 512 tokens, while Sentence-BERT can handle sequences of up to 2048 tokens. This allows Sentence-BERT to better handle long documents, such as articles or books, and provide more accurate embeddings for these types of texts.

In addition, Sentence-BERT can be tuned for specific tasks or industries, which can improve the performance of the embeddings and the classifier. The pre-trained model can be fine-tuned on a smaller set of labeled data, which allows it to learn the specific characteristics of the task or industry, resulting in improved performance.

In summary, Sentence-BERT can improve upon basic BERT by providing more accurate embeddings for text classification, question answering, and named entity recognition tasks. It can handle longer input sequences and can be fine-tuned for specific tasks or industries, resulting in improved performance. Sentence-BERT can be considered an enhanced version of BERT, providing a more powerful and versatile tool for natural language processing tasks.

CLIP

CLIP (Contrastive Language-Image Pre-training) is a training method developed by OpenAI that allows neural networks to learn from both text and images in a supervised way. CLIP is based on the idea of pre-training a neural network on a large dataset of images and associated text, and then fine-tuning it on a smaller dataset of labeled data for a specific task.

The CLIP training process involves training a neural network to predict the text associated with an image, given a set of candidate texts. The network is trained using a contrastive loss function, which maximizes the similarity between the predicted text and the correct text while minimizing the similarity between the predicted text and the incorrect texts. This process allows the network to learn representations of both images and text that are useful for a wide range of natural language processing tasks.

CLIP can be used to join sentences from different fields by training a model on a dataset containing images and text from multiple fields such as news, social media, scientific literature, etc. The model will learn to extract information from text and images and will be able to understand the relationships between the sentences and the images.

Once the model is trained, it can be fine-tuned on a smaller dataset of labeled data for a specific task, such as text classification or question answering. The fine-tuning process allows the model to adapt to the specific characteristics of the task and to provide more accurate predictions.

In conclusion, CLIP is a powerful training method that allows neural networks to learn from both text and images in a supervised way. It can be used to join sentences from different fields by training a model on a dataset containing images and text from multiple fields. The model will learn to extract information from text and images and will be able to understand the relationships between the sentences and the images, providing a more powerful and versatile tool for natural language processing tasks.

Methods

Results Anticipated

Conclusion

Reference

Climent, F. (2018). Ethical Versus Conventional Banking: A Case Study [Number: 7 Publisher: Multidisciplinary Digital Publishing Institute]. *Sustainability*, *10*(7), 2152.

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.

Murè, P., Spallone, M., Mango, F., Marzioni, S., & Bittucci, L. (2021). ESG and reputation: The case of sanctioned Italian banks [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/csr.2047]. *Corporate Social Responsibility and Environmental Management*, *28*(1), 265—277.