Department of Digitization (DIGI)

M.Sc. in Business Administration and Data Science

Natural Language Processing and Text Analytics

Final Examination Project

Evaluating Sustainability Impacts based on Company Tweets

Authors:
Henry Stoll, 141636
Alisa Ilina, 141804
Simon Christensen, 111803
Manuel Escotto Kuhn, 146907

Supervisors:
Raghava Rao Mukkamala
Sine Zambach
Rajani Singh
Weifang Wu

June 15, 2021

Pages: 15

Abstract

In the face of changing customer demands and rigorous stakeholders' expectations for sustainable corporate practices, Environmental, Social and Governmental (ESG) reporting has become increasingly relevant for organizations. This paper uses Twitter data to predict the ESG risk score category of American companies in order to answer the research question of whether there is a correlation between company tweets and their sustainability rating. The dataset used consists of a merge between Twitter handles of 300 American companies (comprising 500 000 tweets in total) and ESG risk scores obtained from Yahoo Finance. After cleaning and tokenization, several NLP vectorizing techniques were compared and piped into four different traditional classifiers. Parallelly, pretrained embeddings were used for prior context in order to apply and compare the performance of different neural language models — RNN, GRU, LSTM and 1D CNN. Lastly, transformers were also applied as a third analysis variant using BERT. Overall, all implemented models performed within a 54% to 64% range of accuracy, which is promising in showing some correlation between tweets and ESG score prediction. However, a tradeoff between complexity of implementation and lack of explainability of results may hinder a widespread adoption of such an analysis at an organizational level. Future research directions can be the use of a larger dataset with a wider range of companies, different pre-processing techniques, training embeddings from scratch, using HAN and LDA.

Keywords: ESG Risk score, Twitter, Embeddings, RNN, GRU, 1D CNN, Vectorization, BERT

Contents

1 Introduction

The social and ethical dimensions of businesses play an increasingly vital role in the sustainable growth of organizations. This concept is commonly referred to in the literature as corporate sustainability, which can be defined as the impact a firm's policies in terms of environmental, social, and economic governance (ESG) have on society (Artiach, Lee, Nelson, & Walker, 2010).According to the United Nations Sustainable Stock Exchange (2015), by 2030, all large companies are expected to report and disclose their ESG practices. The importance of reporting such information has been backed up by previous research, where authors argue that the disclosure of non-accounting information can serve as an indicator of how a firm manages the business risks, and that better ESG scores were even potentially associated with lower business risks (Sharfman & Fernando, 2008).

However, while there is compelling evidence that demonstrates the relevance of ESG ratings for companies, sustainable reporting is still in its infancy and the determinants of a good or bad ESG score have not yet been fully established. On the one hand, as argued by (Elzahar et al., 2015), since ESG information is a type of non-financial disclosure and thus does not follow a standardized reporting format as financial documents do, ESG reporting tends to be rather heterogeneous. On the other hand, while there exist multiple sustainability indices, such as the Dow Jones Sustainability Emerging Markets Index and the FTSE4Good Emerging Indexes, ESG rating agencies normally use different assessment criteria and research methodologies to evaluate companies (Escrig-Olmedo et al., 2019), which further complicates the task of determining which specific factors might influence ESG ratings. While there are multiple factors that might influence the ESG scoring of organizations, little attention has been given to the link between these ratings and social media, specifically in the case of Twitter, which is the most frequently used social media platform of the Fortune 500's (Culnan, McHugh, & Zubillaga, 2010).

With the exponentially increasing volume, velocity, variety and veracity of social media data, as well as sophistication of technologies, Twitter becomes a crucial source of acquiring sustainability impact information. Such dynamic is manifesting the potential for Machine Learning (ML) and Natural Language Processing (NLP) techniques to become the method of critical information extraction from social media. Therefore, based on this premise, the paper will aim to answer the following research question:

Is there a correlation between company tweets and its sustainability rating?

2 Related Work

While there is still no previous research on the specific topic of ESG rating prediction based on Twitter sentiment analysis, other studies have analyzed the predictive power that tweets have for different organizational indicators. For instance, Sprenger2014 analyzed the sentiment of ca. 250.000 stock-related tweets using Naïve Bayesian text classification and established that they can serve as proxies for belief formation and investment behavior.
Speaking specifically about corporate social responsibility (CSR)(Chae & Park, 2018), discovered that textual data, particularly social media, is heavily loaded with CSR-related topics, some of the most popular ones being company strategy, community charity (such as crowdfunding initiatives), philanthropy, and green initiatives, among others. This study showed that companies tend to use social media more as a stakeholder management tool (such as communication of cultural projects) rather than for communication of topics with societal impact (such as equality in the workplace).

In the research by (Yang, Macdonald, & Ounis, 2018), election tweet binary classification task was performed. The authors evaluated the impacts of parameters in embedding training process, such as context windows, dimensionality and number of negative samples. The findings showed that the right choice of embedding model used with CNN led to drastic improvements compared to baseline and random classifiers. In general, numerous papers observed word embeddings followed by deep neural networks to be effective in Twitter related classifications, such as, for instance, in sentiment analysis performed by (Severyn & Moschitti, 2015) or sentiment-specific word embedding Twitter corpus framework proposed by (Collobert & Weston, 2008). However, literature reveals rather few multinomial classification tasks based on Twitter were performed.

3 Conceptual Framework

Twitter is a microblogging service that allows its users to publish short messages of up to 280 characters, popularly known as "tweets" (Sprenger et al., 2014). Tweets differ from regular text as little information gets repeated, thus rendering a string of independent characters. As Tweets are limited to 280 characters, abbreviations, grammar and syntactic errors, and slang are likely to appear which might be unknown to the used pre-trained embeddings, e.g. GloVe. Moreover, hyperlinks, mentions, hashtags and other internet-specific terms are common. Twitter often has noisy data, with texts uncommon to regular corpora (Sosa, 2017)

ESG risk rating measures the extent to which ESG issues are putting company's enterprise value at

risk, particularly the magnitude of the unmanaged ESG absolute risk which is comparable across industries and companies. Exposure to risk, meaning the vulnerability of it, takes into account the industry as it affects the ESG risks a company may face. The metric also includes the management processes in place, such as health and safety systems, employee fatality rate, and others.

4 Methodology

4.1 Data Analysis Process/Workflow

The workflow is presented in the diagram above in Figure 1, with the models sorted in the descending performance order. The tweets of 300 American companies are merged with ESG scores from Yahoo Finance and the ESG Score binned in risk categories. After cleaning and tokenization, the tweets were used to perform multinomial classification per ESG risk categories with accuracy as the key performance evaluation metric. Grid Search was employed to find the best combination of vectorizers and classifiers to serve as a baseline. Learnings from the traditional models were used in selecting Embeddings and then used for comparing different DNN Architectures. Lastly different pretrained BERT Models were used to compare against the results of DNNs.

4.2 Data Acquisition

Twitter handles and usernames were manually obtained for 300 American companies. The `Twitter API v2` was first used to fetch the user ID based on the Twitter handle. Then, tweets were crawled for each company username from May 2021 up to 2019, resulting in at max 3250 tweets per company (due to API limitations) amounting to 500 000 tweets in total. The ESG Risk scores were obtained by web-crawling Yahoo Finance. ESG scores category bins were used as target labels, thus identifying the task as multinomial classification.

4.3 Data Pre-Processing and Normalisation of Tweets

The data was joined with the tweets and split into 50 tweets per row. Any company that had under 50 tweets was discarded. Despite Sustanalytics providing five categories for ESG risk scores, only three bins (low, medium, high) were used, omitting the first category (negligent) since no companies in the dataset depicted negligent ESG risk. Furthermore, the severe risk only contained tweets that amounted to about 3% of the total number of tweets crawled, and thus it was decided to merge with the high category based on the companies ranging very close to the limitation line.

Pre-processing techniques reduce noise and render a consistent transformed dataset further to be used in modelling. Prior to natural language processing of tweets, the text was normalised and pre-processed with the use of tokenisation, lemmatisation and stop words removal. Additionally all unicode characters were converted to there ASCII representation. Tokenisation segmented the tweets into separate tokens, thus extracting relevant features, further the tokens were lower-cased. It is common practice to use deterministic techniques based on RegEx since carefully crafted algorithm can help avoid any ambiguities in text (Klabunde, 2002). Thus, RegEx was utilized for hashtag and username symbols removal. Then applying the processing of `TextVectorizer` of the `Keras` library, it split each sample into substrings (usually words), then recombined the substrings into tokens (usually n-grams). Afterwards, it indexed tokens (associated a unique `integer` value with each token) and finally transformed each sample using this index, into a vector of `integers`. After the initial data pre-processing, the TextVectorizer in the Keras library analyzed the dataset, whereafter it determined the frequency of individual string values, and created a 'vocabulary'. The output is a vocabulary that contains either an unlimited or capped size, depending on the configuration options. If there exist more unique values in the input than the maximum parsed vocabulary size, the most frequent terms will be used to create the vocabulary.

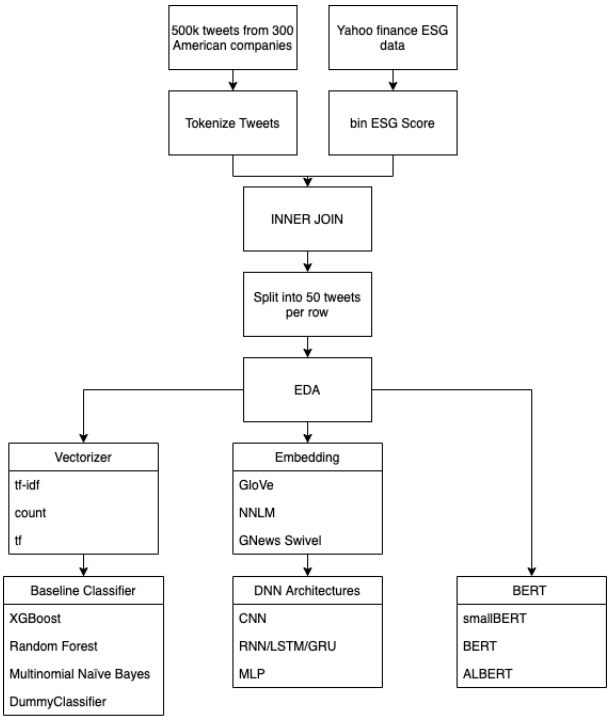

Figure 1: Process Diagram

4.4 Dataset Description

Exploratory data analysis reveals the dataset to be quite vast, yet the distribution of ESG Risk scores depicts the majority of companies scoring low and medium ESG Risk, making high and severe categories relatively under-representative (Figure 4). The dataset only contains recent ESG rankings and does not account for historical ESG labels. WordCloud was implemented to visualise most common words used (Figure 5). The average tweet length is 170 characters (Figure 6).

The industry scoring highest in ESG Risk is the Energy, possibly due to its wide exposure to environmental risks, as it is one of the three categories comprising ESG Risk score (Figure 3). Sentiment analysis using VADER was performed to get an overview of emotional context of Tweets. Identifying companies with least positive tweets (lowest compound score), NewsCorp displayed lowest sentiment score, possibly since the tweets are associated with news events content likely to be rather impassive. Sentiment analysis shows that among most common negative words used are "apologise, sorry, issue, help, inconvenience, frustration", while positive words include "thanks, learn, experience, information, team".

The most frequently used words (unigrams) for each of the three categories, (low, medium, high) can be seen in Figure 2. The graph clearly indicates that there is no visual difference between the different categories according to the figure. 15 out of the top 20 terms across the different categories are the same, which may indicate it being difficult to differentiate between them. Furthermore, it shows that the categories, low, medium and high have respectively 169.050, 222.267 and 114.906 unique terms while there in total are 149.500, 228.850 and 104.250 tweets in the above categories.

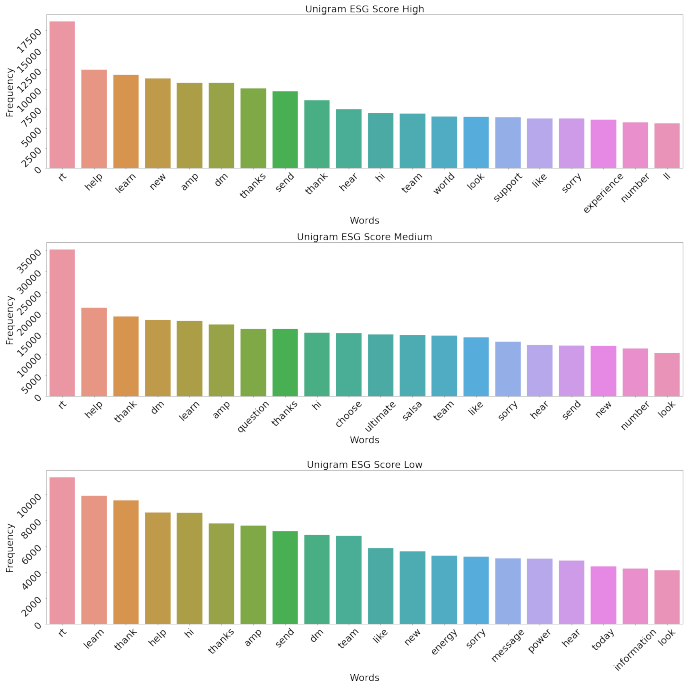

Figure 2: Most frequently used words per label

4.4.1 Traditional Classifiers

Prior to the classifiers, several NLP vectorizing techniques were compared. The best vectorizer was selected with Grid Search and piped into traditional classifiers to establish baseline metrics.

Count Vectorization leverages frequencies in which a token appears in a tweet. The CountVec-

torizer generates a sparse document term matrix with each cell representing the count of terms in a document given a vocabulary. To demonstrate, given a vocabulary `['exams','are','stressful']` and a sentence `'Exam is life, but life is somewhat stressful'` Would result in the following vector: `[1, 0, 2].` Sklearn's implementation defines an n-gram range for their implementation. Without supplying the class with an n-gram range (m, n), it follows a Bag of Words model.

TFIDFVectorization computes a sparse model which defines word meanings by computing term frequencies (tf) and inverse document frequency (idf). tf assists in identifying the frequency of word t in the document d, as the `CountVectorizer` described above. A common technique is to lower the term frequency to avoid skewness, by taking $\log_{10}$ which results in the following calculation:

$$tf_{t,d} = \begin{cases} 1 + \log_{10}, & \text{if count(t,d)} > 0 \\ 0 & \text{otherwise} \end{cases}$$

The idf is computed by $idf_t = \log_{10}(\frac{N}{df_t})$, where N is the total number of documents and df_t is the number of documents which contain the term t. It is used to allocate higher weight to words in only a few documents. The difference in performance of $tf - idf$ vectorization compared to count vectorization will be analysed in section XX (results) and XX. `max_features` will be considered in order to optimise the train time. In order to obtain a high $tf - idf$ weight for a given term, the term needs to contain a high term frequency (tf) in a given document (local parameter) and a low document frequency of the term across all documents (global parameter). **N-grams** also generate a document term matrix, yet the count here is the combinations of adjacent tokens in length n. A smaller n might not explain enough information, while bigger n will overload the information on features. *The best performing vectorizer was evaluated and piped into the following classifiers:*

Dummy Classifier makes predictions using simple rules. With the strategy `stratified` it generates predictions with respect to the training set's class distribution in order to compensate for the small class imbalance in the dataset. It is used to create a simple baseline to compare with other classifiers. **Multinomial Naïve Bayes Classifier** is a frequency-based model designed to determine the number of times a specific word or term occurs in a document. It assumes that all features have same importance and that each feature is independent of others, resulting in it not accounting for the order of tokens. Being one of the key classifiers used for text classification with discrete features and given its easy implementation and efficiency, it is well suited for the purpose of this research. (Wiratama & Rusli, 2019). **Random Forest** addresses the risk of overfitting that

arises from making the classifiers learn from too many features and from potentially noisy data (as is the case of tweets). The individual decision trees that comprise the random forest may perform well on prediction, but they could eventually overfit a portion of the dataset. An averaging of the results is an effective alternative to overcome this issue (Müller & Guido, 2016). **XGBoost** is a scalable machine learning system for tree boosting. It achieves optimal results by training each new instance with the aim to emphasize the training instances previously mis-modeled for better classification results (Li, Zhang, Wang, & Wang, 2020). The algorithm can be best described as a combination of classification and regression trees, but with a re-defined objetive function in terms of complexity of the trees and training loss in order to mitigate the risk of overfitting the data, which allows it to be a highly robust and accurate model.

4.4.2 Neural Language Model Architectures

Generally, neural networks have been widely adopted for NLP tasks. Neural language models use neural networks as a classifier computing probabilities. They do not require smoothing and are outstanding in generalisation of context and handling histories, often outperforming n-gram models, thus are often expected to reach high performance scores. However, the training time, costs and complexity need to be considered. The prior context in neural models is represented with the use of embeddings. Pretrained or 'learned from scratch' embeddings may be used, yet due to time constrains, only pretrained embeddings were used and fine-tuned during model training.

Vector Semantics and Embeddings

Vector semantics instantiate distributional hypothesis on meaning similarity of words by learning embeddings from their distributions in tweets. Vector semantics in general aid in representation of a word in multidimensional semantic space crafted by distributions of neighbouring words. Word and document similarities are most commonly calculated with the normalised dot product. Embeddings are short dense vectors, which outperform sparse vectors like those generated by n-grams due to smaller parameter space aiding in avoiding overfitting and generalisation. Both static and contextualised (for instance, BERT) embeddings were used. They are a part of representation learning which has been featured prominently in recent NLP research. Word embeddings based on neural networks have recently been widely adopted as an alternative to traditional vectorizers, such as `tf-idf` (Bengio et al., 2003). The below introduced embeddings (both token and sentence types) were compared against each other by feeding them into a fully connected layer with 128 units and then into a output layer with softmax.

GloVe is an unsupervised algorithm which is used to generate vector representations of words. It uses probability ratios of a co-occurrence matrix and is able to display linear substructures of the word vector space (Jurafsky & Martin, 2002). Pretrained embeddings from the 6B uncased tokens corpus were used which provides et al. 100 and 300 dimensional representation of each word. GloVe was implemented by transforming each token of a tweet to an n-dimensional embedding, creating a sequence of embeddings to be used in sequential models like 1D CNNs, RNNs and a combination of them both. Even though GloVe is pretrained, the embedding is still trainable and was fine-tuned during fitting. The embedding layer additionally created a zero mask, improving the fit performance as it allowed following layers to ignore any zero vectors. The **Submatrix-wise Vector Embedding Learner (Swivel)** is a count-based model trained on the English Google News 130GB dataset that generates low-dimensional feature embeddings based on a co-occurrence matrix. "Swivel uses stochastic gradient descent to perform a weighted approximate matrix factorization, ultimately arriving at embeddings that reconstruct the point-wise mutual information (PMI) between each row and column feature" (Shazeer et al., 2016). Word embeddings are combined into 20 dimensional sentence embedding using the weighted sum divided by the square root of the sum of the squares of the weights (`sqrtn`). **Neural Network Language Model (NNLM)** is another embedding model trained on Google news that uses neural networks consisting of three layers to learn the word embeddings. This model uses various previous words as context for the target word prediction, ultimately outputting a softmax layer that can translate vectors to probabilities. The input layer is a concatenation of all context words to capture order information, the hidden layer projects the input layer to another vector space, and the output layer consists of units, each representing a word in the vocabulary (Zheng, Shi, Guo, Li, & Zhu, 2017). 50, 128 and 128-normalised dimensional NNLM sentence embeddings were used (Figure 9).

Deep Neural Network Architectures

"Language is a sequence that unfolds in time", Jurafsky and Martin (2002, pg. 172). Approaches which do not account for a time dimension assume simultaneous access to inputs, making it troublesome to address temporal nature of text. Many language tasks require access to arbitrarily distance information. Recurrent architectures and transformer architectures address these issues and allow to capture temporarity of text, which will be further analysed through the prism of probabilistic models which assign conditional probability to the next token. The traditional n-gram or feed-forward networks are facing constraints of the Markov assumption. However, the sequential and transformer architectures manage to use greater contexts weakening this equation.

Recurrent Neural Network (RNN) is a sequential algorithm with cycled connections and dependence of values upon previous inputs. It introduces recurrent neurons which maintain memory state from the past. Thus, the outputs of units are based both on current input as well as previous values of the hidden layer. RNNs are trained with backpropogation through time and commonly used for sequence classification. However, they face risks of unstable gradients and are challenging to train. Thus, more complex architectures such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) were implemented as those contain gates to control information flows, and are expected to perform better. **LSTM**, a more complex variant of RNN, is able to extract long-term dependencies in word sequences with the use of gates to control information flow within the recurrent cell. **GRU** is simpler than LSTM, yet still handles longer sequences than RNN. It combines forget and input gates and returns a state vector at each input (Geron & Géron, 2017). All RNNs were implemented by feeding the sequence output of the embedding layer into one or multiple recurrent layers with hidden-to-hidden connection scheme, connected to a hidden-to-output layer, condensing the sequence for the following softmax output layer (see figure X). Since GRUs have the fastest training time, they were implemented first to find the best GloVe embedding dimension. Secondly, their performance was compared against LSTMs and simple RNNs. Additionally, the impact of the number of hidden layers on performance was considered.

Convolutional Neural Networks have revolutionised various areas of machine learning, carving their way to becoming a building block for NLP Shen, Min, Li, and Carin (2020). Convolution, the mapping of a filter to an input to extract features, results in a feature map, which indicates patterns of detected features in an input. Word embeddings together with CNNs particularly have received wide attention in text classification tasks for Twitter (Yang et al., 2018). Specifically when used with pre-trained word embeddings, CNNs show outstanding performance empirically outperforming traditional classifiers. According to Goldberg (2016), the easy integration of pre-trained embeddings, ability to pick crucial tokens or sequences of tokens in a manner which is invariant to the inputs, makes CNNs particularly strong in their performance. CNNs are also especially useful in classification of problems where strong local signals of class, which may occur in various input parts, are expected. Such observation is highly relevant for the given task of this research, since the goal is to identify whether certain sequences of words, despite their position, may indicate class membership. CNNs peculiarity is the ability to learn such local indicators in varying locations. Literature suggests the following architecture, which has further been implemented in the workflow: (1) word embedding, (2) convolutional model which learnt to extract important features

from word embedded tweets, (3) fully connected model which interprets the features and performs a classification. In a study by Conneau, Schwenk, Cun, and Barrault (2017), the authors prove that deeper architectural approach for NLP tasks is often beneficial. To perform ESG classification the authors used a CNN architecture inspired by Kim (2014) (see Figure 7). GloVe embeddings were again used to generate the inputs sequences. Those was fed into multiple parallel convolutions with kernel sizes of $(2, 4)$ each with 64 filters. A convolutional dropout of 0.5 was applied to each feature map, encouraging better defined features. In accordance with (Geron & Géron, 2017) MaxPooling was applied as it outperforms other types of pooling in NLP tasks. The pool size was set to 2 to reduce features, and a stride of 1 was used to examine every combination of words. Global max pooling layer reduced the sequence data to a pooled output. All layers then were concatenated and fed into one fully connected layer. ReLU activation function was used due to its empirically superior performance in CNNs (Glorot, Bordes, & Bengio, 2011). In the output layer, softmax was used to perform classification (Figure 7). L2 regularisation ($\lambda = 0.0001$) was used to add penalty to the complexity and avoid overfitting. A dropout rate of 0.5 was added at the before the output layer, as recommended by (Geron & Géron, 2017).

4.4.3 Transformer Architecture and Bert

Despite the fact that sequential architectures assist in retaining previous information, the problem of relevant information and efficient training still remains. Transformers were developed to tackle such shortcomings. In addition to standard network components, transformers use self-attention layers, which allow to capture information from vast context, mitigating the necessity to transfer such information through recurrent connections. Self attention layers map sequences of input vectors $(x_1, \ldots, x_n)$ to outputs $(y_1, \ldots, y_n)$ of the same length, each item's computation is independent of others. This architecture can compare the item to others in terms of relevance in the given context. The comparison in its simplest essence is computed with score $(x_i, x_j) = x_i \cdot x_j$.
The larger the output, the more similarity vectors share. Then, scores are normalised with Softmax to output a weights vector $\alpha_i j$. Given proportional scores in α, output value is obtained by summing the inputs already seen by the architecture, weighted with corresponding α values. $y_i = \sum_{j<=i} \alpha_{ij} x_j$ Thus, the attention-based method involves the following steps: comparison of the relevant items, normalisation to generate probability distribution, and computation of weighted sum of the obtained distribution (Jurafsky & Martin, 2002). Transformers, according to Vaswani et al. (2017) is the first model to rely fully on self-attention while generating input and output representations avoiding RNN or convolutions.

Bidirectional Encoder Representation from Transformers (BERT) is a deeply bidirectional pre-trained model architecture, meaning it interprets the text from both sides, as opposed to traditional sequential methods. Transformer, the underlying mechanism of BERT, interprets contextual relationship with the use of encoders and decoders. Encoders capture the input — the sequence of tokens, while decoders build predictions (Mishra, Rajnish, & Kumar, 2020). Several BERT types were implemented. **A Lite Bert (ALBERT)** is an advancement on the performance of the conventional BERT by allocating the model's capacity more efficiently. In contrast to the normal BERT, the input-level embeddings learn context-independent representations (Soricut & Lan, 2019). Similarly, the **small BERT** models are based on the main BERT architecture only with a smaller number L of layers (L2 and L4 in this case) together with a smaller hidden size H and a corresponding reduced number A of attention heads (Turc, Chang, Lee, & Toutanova, 2019).

5 Results

Overall, all models implemented performed quite similarly, in the range of 54% to 64% accuracy, exhibiting some correlation. Hence, even simple models result in rather similar accuracy scores. However, given the complex nature of the multinomial classification, even small differences are important and provide distinct insights into architectural differences.

Traditional Classifiers

The tf—idf vectorizer outperformed the count vectorizer by 1.24% and the tf vectorizer by 2.31% when piped into a Random Forest Classifier. This means that infrequent tokens are relevant in ESG Risk score prediction, likely since sustainability related patterns of tokens are uncommon in tweets. Unigrams performed best (0.562%) compared to bigrams (0.543%) and trigrams (0.517%), meaning that single tokens in the tweets are crucial, and that sustainability is likely more signalled through single words. The Dummy Classifier establishes 37% random guess benchmark for the given multinomial classification, meaning classifiers scoring higher than that are better than guessing at random. XGBoost reached 60.7% accuracy outperforming other traditional classifiers (see Table 1). This could be due to XGBoost being more robust towards class imbalance and capable of handling noisy data, such as tweets.

Classifier	Accuracy
$\text{DummyClassifier}_{s=\text{stratified}}$	0.370
MultinomialNB	0.576
RandomForestClassifier	0.562
XGBoostClassifier	0.607

Table 1: Traditional Classifiers as Baseline

Neural Network Architectures

Sentence embeddings achieve a higher accuracy when compared to the traditional classifiers, scoring second best overall when using nnlm-en with 50d. This is due to sentence embeddings providing a low dimensional representation of all the tweets and ignoring information regarding word order. However, to compare against the performance of sequential models and examine whether sequences improve performance, GloVe embeddings fed into sequential models were investigated. (see Table 2)

Model		Metrics
Embedding	Dimensions	Accuracy
nnlm-en	50	0.641
nnlm-en	128	0.579
nnlm-en normalized	128	0.637
GNews Swivel	20	0.631
GloVe	300	0.566

Table 2: Comparing Sentence Embeddings

300- and 100d embedding were considered, 300d worked better due to capturing more data (see Table 3). Comparing the three recurrent models against each other, the SimpleRNN performed worst whilst the LSTM achieved a lower accuracy compared to GRUs, possibly because the model is not benefiting from having long term memory states. Different number of hidden layers were considered, stacking GRUs did not improve results (see Figure 8). These findings show that in order to predict a ESG score, positional information is not necessarily needed, which makes sense considering the average length of a tweet.

Model			Metrics
Embedding	hidden layers	units	Accuracy
GloVe 300d	GRU	300	0.611
GloVe 100d	GRU	100	0.602
GloVe 300d	GRU	50	0.576
GloVe 300d	LSTM	50	0.564
GloVe 300d	SimpleRNN	50	0.561
GloVe 100d	2x GRU	100	0.577
GloVe 100d	3x GRU	100	0.543
Glove 300d	CNN [2, 3, 4]	64	0.644

Table 3: Sequential Architectures

CNN with 300d, global max pooling, scored highest among all models. That confirms the fact that CNNs are work well in the given scenario as they are able to classify classes with strong local signals, despite the specific position.

Transformers

Transformers did not manage to reach higher performance, meaning that the self-attention layer and context are not crucial. SmallBERT scored highest (see Table 4). Thus, sustainability score is

better predicted not by the context but by singular words, disregarding positioning of the words, consistent with the above-mentioned results of traditional classifiers and neural language models.

6 Discussion

With an accuracy of the best performing model of 0.644%, it is determined that there exists a correlation between company tweets and its sustainability rating. Looking at the dummy classifier compared to the best performing model, it clearly shows a sizeable percentage difference. This does hint that there are features that are independent and thus support the argument for a correlation. On the other hand, it is arguable that despite the correlation there is no substantial usage of a model that classifies 0.644% of the labels correctly. It must though be weighted against how 'wrong' a classification is, meaning to which extent a business case would need exact results.

	Model			Metrics
Model	Layers	Hidden	Heads	Accuracy
smallBERT	2	128	2	0.531
smallBERT	6	256	4	0.609
smallBERT	10	512	8	0.624
BERT uncased	12	768	12	0.601
AlBERT base	12	768	12	0.556

Table 4: Comparison of BERT based architectures

While the models applied showed an acceptable accuracy for predicting ESG categories based on Twitter data, a trade-off between complexity and explainability becomes relevant to discuss. On the one hand, multiple studies (such as Chae2018) have demonstrated that text data, specifically Twitter data, contains a wide topical landscape and discusses multiple CSR dimensions of an organization, which could hold great potential for the endeavor of predicting ESG scores based on this data. However, on the other hand, this paper has demonstrated that, while promising, Twitter data alone is still not sufficient to predict the ESG risk scores of a company. Likewise, a further constraint for the applicability of the results within an organizational context is the workforce skills necessary to implement such a project. "To take advantage, CSR researchers and practitioners need to be familiar with computational data collection techniques such as web crawling and APIs" (Chae & Park, 2018). Therefore, the constraints of limited ESG scores explainability from Twitter data and the complexity of implementing such an analysis are two relevant factors companies should take into account when deciding whether to perform such type of analysis.

Contribution and Implications for Research. This paper provides insights into the missing research gap on correlation and impacts of Tweets on ESG Risk scores. The potential linkage between these indicators could affect the way sustainability is signalled, evaluated and perceived. The findings depict the need to consider other sustainability signals for a more accurate decision, (potentially, other social media sources, such as LinkedIn, or for even more direct influence — ESG reports) in tandem with Tweets, as they might not deem sufficient for accurate prediction of sustainability performance in some business contexts given the tradeoff of misclassification.

Learning Reflections. It has come to attention during the workings of this paper that making a significant and in-depth exploratory data analysis has its grounds. Not only for ensuring a thorough understanding of the dataset, but also the need for specific train-test splits, imbalances and so forth. Significant time was spent using a random test-split, which brought bias to the models. Retraining a significant amount of models and finding the right hyper-parameters is computationally heavy and does require substantial amount of time which could have been avoided by understanding the data-set better from the start.

Limitations of the dataset/work. As described in subsection 4.4, there is a significant reduction in the number of categories from five to three, where both negligent and severe were dropped. This of course removes the possibility of predicting the two dropped categories. Further data gathering of tweets of companies that range in both these categories would allow a more diverse and accurate prediction of the entire ESG Risk Ratings. Furthermore, the data was limited to 300 companies with their corresponding tweets. Internally, the Twitter API has a limit of 500.000 acquired tweets per month, which sets a limitation to the amount of data being able to be gathered. Having a higher tweets limit would have allowed us to specifically balance the dataset better, allowing for a more accurate and balanced set and thus presumably also allowing for better, more accurate and wider prediction range. Moreover, GloVe was not trained from scratch, only fine-tuned, also posing a limitation. Thus, the limitations of the data set clearly impacted both performance and overall prediction level, which could have been avoided by either balancing the initial 300 companies more or by having a larger Twitter API rate.

It is to be noted that the possibility of continuously using the same model for predicting the ESG Risk Level of a company based on their tweets may with time yield worse results. As stated in section 3, tweets are limited to 280 characters, which causes abbreviations, grammar and syntactic errors along with the possibility of slang. Furthermore, the important factors of both the

environment, social and governmental aspects may change over time, making the trained models potentially perform worse due to the underlying patterns changing. On the other hand, time will likely change a companies ESG score. With the model being trained on tweets based on ESG Risk Scores and only partially relates to a given company, re-evaluating the companies tweets after a period would reflect the current sustainability level.

Future Work Currently, as shown in section 5, there only exists a limited a correlation between a company's tweets and their ESG Risk Score. Due to scarce research, trying to link companies' social media profiles to their ESG Scores, it makes it a subject worth of further research. The paramount factor for potential better results is to gather a substantial amount of additional tweets from a wider range of companies. Furthermore, an initial study could be made to discover to what extent the two discarded categories, negligent and severe, each of which is in the extremes of the scale, have significant features that would allow for better classification. Other approaches could also be taken, trying out different versions of GloVe such as the Twitter embedding (and possibly training embeddings from scratch) or trying to implement a Hierachical Attention Networks (HAN). Latent Dirichlet Allocation (LDA) topic modelling could be performed to investigate abstract topics in tweets and their links to sustainability. Finally, different data-preprocessing techniques and different combinations of the number of tweets per data row could be adjusted in order to determine the significance in relation to the results.

7 Conclusion

Despite the existence of multiple sustainability indices, there is still a need for efficiently being able to evaluate to which extent a company delivers on both a environmental, social and governmental level. With Twitter being the most widely used communication platform of the Fortune 500's, this paper tried to find a correlation between companies' tweets and their respective sustainability rating. Cleaned and tokenized tweets from 300 different American companies associated with their respective ESG Risk Scores were used and despite the Exploratory Data Analysis not indicating a clear difference between the different ESG Risk tiers, the tweets were used to perform multinomial classification per ESG risk categories. Grid Search was implemented to find the best combination of vectorizers and classifiers to serve as a baseline. Furthermore, learnings from the traditional models were used in selecting Embeddings and then implemented in order to compare different DNN Architectures. Lastly different pretrained BERT Models were used to compare against the results of DNNs. The results did reflect the initial findings of the EDA, where there only to some

extent exist a correlation between a companies tweets and their respective ESG scores. To which extent a business is willing to adopt a model that predicts 64% of the companies' depends on the given business case, but with little research having been done in the subject of linking social media accounts to sustainability indices, there is plenty of room to improve and expand.

8 Appendix

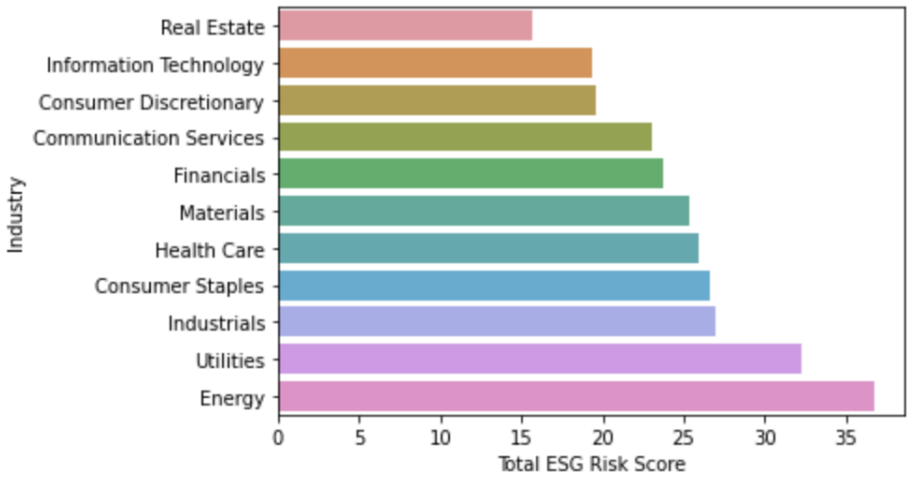

Figure 3: Industries

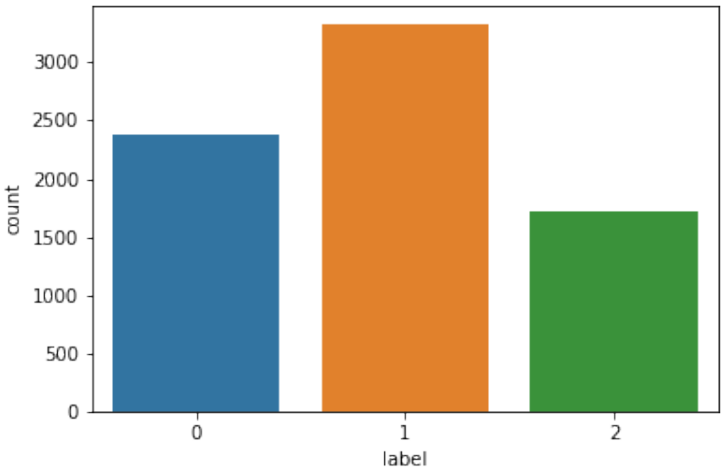

Figure 4: Industries

Figure 5: Words Cloud

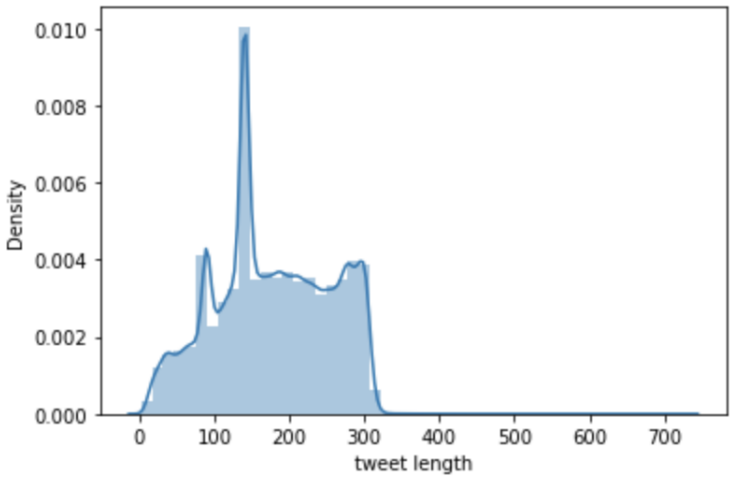

Figure 6: Length of Tweets

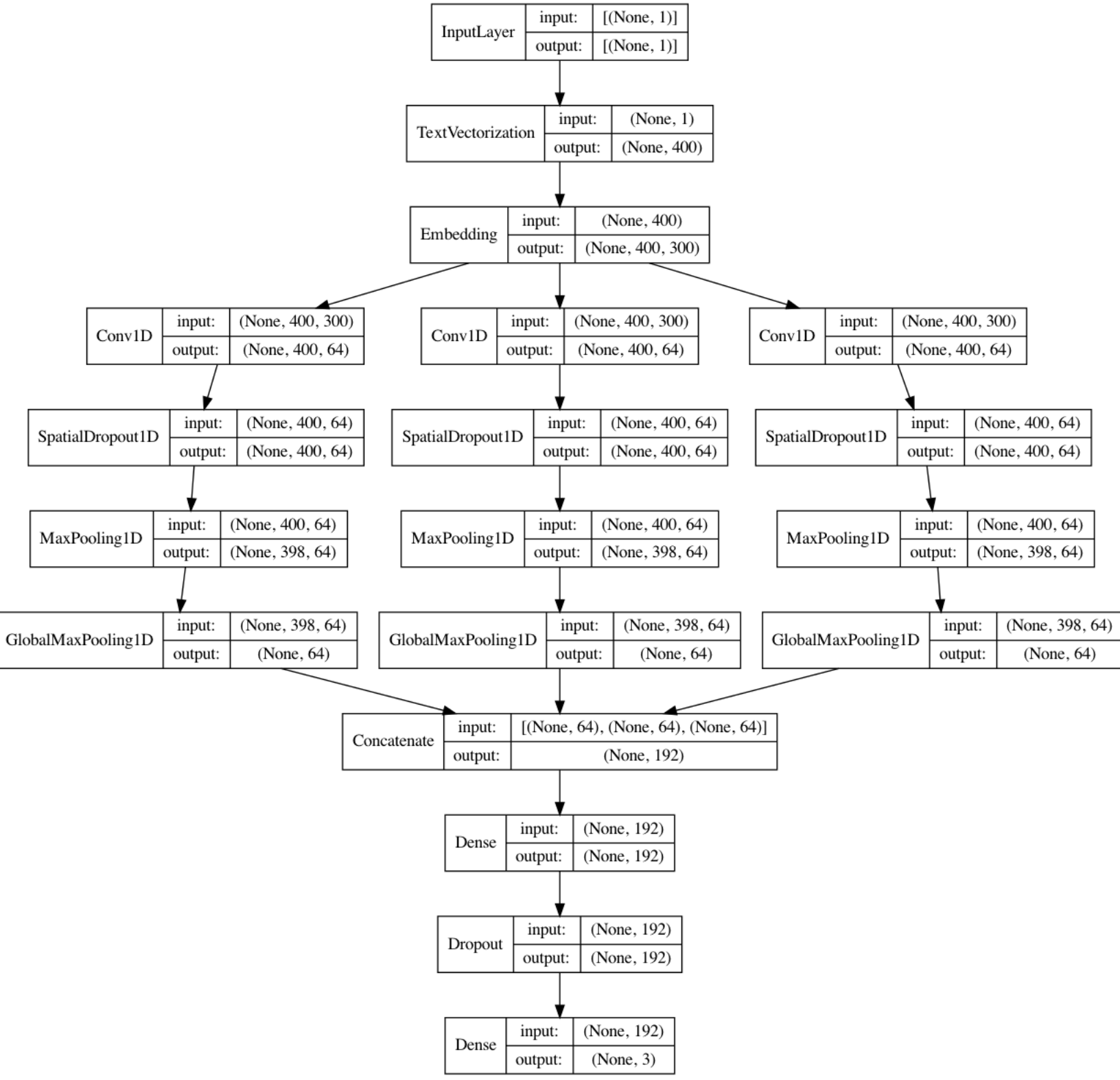

Figure 7: CNN Model with parallel filter sizes of 2, 3 and 4 and GlobalMaxPooling followed by a fully connected layer and dropout

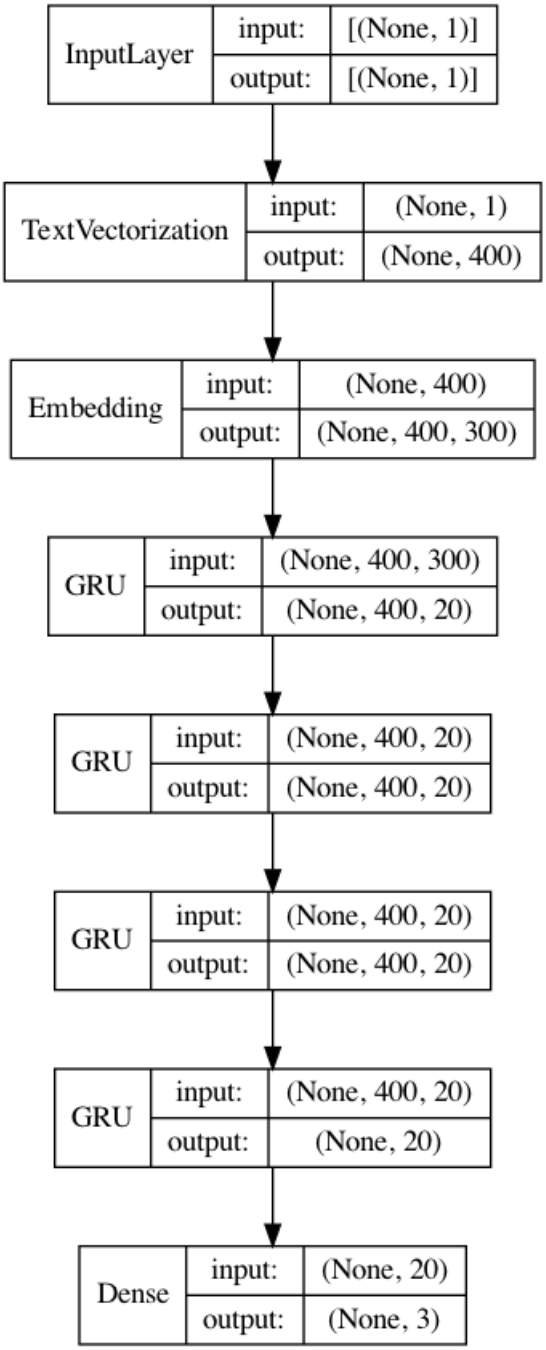

Figure 8: RNN 3x GRU Model with three hidden-to-hidden layers followed by one output-to-hidden layer

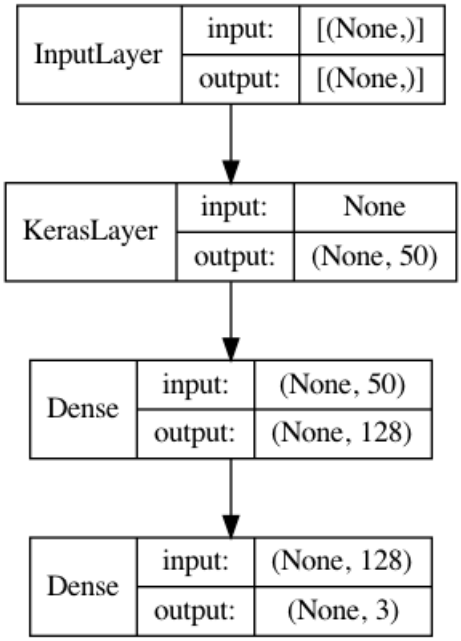

Figure 9: Model using Tensorflow Hub nnlm-en dataset with 50 dimensions followed by a fully connected layer

References

Artiach, T., Lee, D., Nelson, D., & Walker, J. (2010, mar). The determinants of corporate sustainability performance. *Accounting & Finance*, *50*(1). doi: 10.1111/j.1467-629X.2009.00315.x

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A Neural Probabilistic Language Model. In *Journal of machine learning research* (Vol. 3). doi: 10.1162/153244303322533223

Chae, B., & Park, E. (2018, jun). Corporate Social Responsibility (CSR): A Survey of Topics and Trends Using Twitter Data and Topic Modeling. *Sustainability*, *10*(7). doi: 10.3390/su10072231

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on machine learning.*

Conneau, A., Schwenk, H., Cun, Y. L., & Barrault, L. (2017). Very deep convolutional networks for text classification. In *15th conference of the european chapter of the association for computational linguistics, eacl 2017 - proceedings of conference* (Vol. 1). doi: 10.18653/v1/e17-1104

Culnan, M., McHugh, P., & Zubillaga, J. (2010). ow Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. *MIS Quarterly Executive*, 243–259.

Elzahar, H., Hussainey, K., Mazzi, F., & Tsalavoutas, I. (2015, may). Economic consequences of key performance indicators' disclosure quality. *International Review of Financial Analysis*, *39*. doi: 10.1016/j.irfa.2015.03.005

Escrig-Olmedo, E., Fernández-Izquierdo, M., Ferrero-Ferrero, I., Rivera-Lirio, J., & Muñoz-Torres, M. (2019, feb). Rating the Raters: Evaluating how ESG Rating Agencies Integrate Sustainability Principles. *Sustainability*, *11*(3). doi: 10.3390/su11030915

Geron, A., & Géron, A. (2017). *Hands-On Machine Learing With Scikit-Learn & Tensor Flow.*

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Journal of machine learning research* (Vol. 15).

Goldberg, Y. (2016). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, *57*. doi: 10.1613/jair.4992

Jurafsky, D., & Martin, J. (2002). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Zeitschrift fur Sprachwissenschaft*, *21*(1). doi: 10.1515/zfsw.2002.21.1.134

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Emnlp 2014 - 2014 conference on empirical methods in natural language processing, proceedings of the conference.* doi: 10.3115/v1/d14-1181

Klabunde, R. (2002). Daniel Jurafsky/James H. Martin: Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. *Zeitschrift fur Sprachwissenschaft*, *21*(1). doi: 10.1515/zfsw.2002.21.1.134

Li, Z., Zhang, Q., Wang, Y., & Wang, S. (2020, jul). Social Media Rumor Refuter Feature Analysis and Crowd Identification Based on XGBoost and NLP. *Applied Sciences*, *10*(14). doi: 10.3390/app10144711

Mishra, P., Rajnish, R., & Kumar, P. (2020). Sentiment analysis by novel hybrid method be-cnn using convolutional neural network and bert. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(4). doi: 10.30534/IJATCSE/2020/165942020

Müller, A., & Guido, S. (2016). *Introduction to Machine Learning with Python.* O'Reilly Media, Inc.

Severyn, A., & Moschitti, A. (2015). UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification.. doi: 10.18653/v1/s15-2079

Sharfman, M. P., & Fernando, C. S. (2008, jun). Environmental risk management and the cost of capital. *Strategic Management Journal*, *29*(6). doi: 10.1002/smj.678

Shazeer, N., Doherty, R., Evans, C., & Waterson, C. (2016). Swivel: Improving Embeddings by Noticing What's Missing. *ArXiv*.

Shen, D., Min, M. R., Li, Y., & Carin, L. (2020). Learning context-aware convolutional filters for text processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing, emnlp 2018.*

Soricut, R., & Lan, Z. (2019). *ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations.* Retrieved 2021-06-14, from `https://ai.googleblog.com/2019/12/albert-lite-bert-for-self-supervised.html`

Sosa, P. M. (2017). Twitter Sentiment Analysis using combined LSTM-CNN Models. *Academia.edu*.

Sprenger, T. O., Tumasjan, A., Sandner, P. G., & Welpe, I. M. (2014, nov). Tweets and Trades: the Information Content of Stock Microblogs. *European Financial Management*, *20*(5). doi: 10.1111/j.1468-036X.2013.12007.x

Turc, I., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. *arXiv*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 2017-Decem).

Wiratama, G. P., & Rusli, A. (2019, oct). Sentiment Analysis of Application User Feedback in

Bahasa Indonesia Using Multinomial Naive Bayes. In *2019 5th international conference on new media studies (conmedia)*. IEEE. doi: 10.1109/CONMEDIA46929.2019.8981850

Yang, X., Macdonald, C., & Ounis, I. (2018). Using word embeddings in Twitter election classification. *Information Retrieval Journal*, *21*(2-3). doi: 10.1007/s10791-017-9319-5

Zheng, Y., Shi, Y., Guo, K., Li, W., & Zhu, L. (2017, jul). Enhanced word embedding with multiple prototypes. In *2017 4th international conference on industrial economics system and industrial security engineering (ieis)*. IEEE. doi: 10.1109/IEIS.2017.8078651