ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases

Mariya Pavlova
Student Number: 170703132
Supervisor: Dr. Julian Hough
MSc Artificial Intelligence

Abstract—**In recent years, investors and regulators have pushed Environmental, Social and Governance (ESG) investing to the mainstream. This has led to the rise of ESG scores as a way of evaluating an investment's credentials as socially responsible. While demand for ESG ratings is high, their quality varies wildly. Quantitative techniques can be applied to improve, automate, and standardise ESG rankings. A key reason for the limited research in this field seems to be the lack of appropriate datasets. To support research on ESG investing, this paper pioneers a new dataset, the ESG-FTSE dataset comprising of publicly available news articles. This dataset is used to explore the application of natural language processing (NLP) techniques to identify ESG relevance. The paper achieves benchmark classification performance (88% accuracy and 79% F1 score) on prediction of ESG relevance from the dataset. In addition, it successfully applies topic modelling techniques to derive ESG topics and then use them for classification of unseen text. The results demonstrate that ESG-FTSE can be used to create accurate ESG predictions from even low volume data. In consequence, the paper proposes expanding the dataset to include all FTSE 100 constituents, more news publication sources, languages and other data types. Last, the paper suggests using the dataset to create a semi-automated ESG scoring system.**

Keywords—ESG labels, news article, text classification, topic modelling, dataset, small data

I. Introduction

Financial markets have been going through a seismic shift with the rise of ESG investing. The ascent of sustainable investing has also boosted the proliferation of ESG measurement and reporting metrics. Their different requirement and quality have been adding cost, confusion, risk, and complexity to investors. According to recent research, poor data quality is one of the biggest obstacles in the path of ESG investing (Murray, 2021). This, in turn, has prompted concerns over "greenwashing" – i.e., that some investments are not as sustainable as they claim to be. ESG ratings are the most widely used metric. Yet, they have been scrutinised by regulators and investors because of their questionable quality. The paper argues that the limitations of the research methods used to generate ESG rankings is one of the main barriers to accuracy. Even though Artificial Intelligence (AI) techniques can drastically improve the accuracy of ESG metrics and thus of ESG investing, research in this domain is limited. To encourage it, this study takes the novel and challenging approach of creating the first dataset with ESG relevance labels – the ESG-FTSE dataset. In this way, it provides a backbone for future research in this area. Alternative data, such as news articles, have been shown to be a powerful tool for evaluating stock market performance and investment opportunities. By building a dataset that consists of publicly available news articles about FTSE 100 constituents, the paper proves that such data can also be used for assessing ESG credentials. It also demonstrates that NLP techniques can be applied to the new domain of sustainability research. It reveals that there is a small data problem associated with ESG data, i.e., scarce but relevant data. In addition, the paper shows that it is feasible to make accurate ESG predictions by applying both supervised and unsupervised learning methods on the dataset. Last, it explores different avenues for further work. The current study proposes adding all FTSE 100 companies and applying small and wide data techniques to expand the dataset. The research also suggests adding more data types and news publications in different languages to ESG-FTSE. Last, this research explores the possibility of using the dataset to develop a semi-automated ESG scoring system.

II. Hypothesis

This paper recognises the need for accurate and consistent ESG metrics to evaluate an organisation's sustainability standing. It takes a novel approach to addressing the issue by creating the very first ESG dataset. By building it entirely from publicly available news articles, it argues that this dataset would foster research in this subdomain. In addition, the paper takes the view that such alternative data influence not just financial performance but also a company's credentials, thus making it essential to ESG analysis. Second, this project sets out to describe the dataset. Third, it seeks to resolve the small data problem associated with ESG data by identifying whether accurate ESG predictions and topics can be learned from a small ESG dataset, such as ESG-FTSE. To prove the above hypothesis, the current study seeks to demonstrate whether supervised and unsupervised learning approaches can be applied to ESG-FTSE to accurately predict ESG relevance and topics.

III. Data collection of the ESG Corpus

A. Background

Notion of ESG investing. ESG investing is a hypernym for investments that seek positive returns and long-term impact on society, environment, and the business. Environmental criteria may consider an organisation's pollution, waste, energy use, natural resource conservation, carbon footprint, and treatment of animals. The case the miner BHP damaging aboriginal sites, which prompted an inquiry in the Australian parliament, is an example. Social criteria examine a company's management and its relationships with employees, customers, suppliers, and the communities where it operates. The case of wages and conditions of workers in Leicester garment factory, which lead to retailers, reconsidering their purchasing policies, is an example. Governance looks at a company's leadership, executive pay, audits, internal controls, lawsuits, and shareholder rights. [1] The case of increases paid to AstraZeneca investors, which sparked a rejection by shareholders, is an example.

Previously, ESG investing represented a niche area of financial markets. With regulators and investors realising the financial materiality of ESG risks, these financial products have been pushed to the mainstream. Soaring demand for such products has pushed global assets under management of net-zero funds to $43tn. This means half of the global pool of funds is now linked to climate change goals (Mooney, 2021). Their rise in popularity has led to investors seeking more information on sustainability risks. This has given rise to a wide range of initiatives aimed at defining ESG disclosure standards, investment and measurement principles, and metrics (Murray, 2021). ESG scores have become the most widely used metric to measure a company's credentials.

While they are high in demand, the same cannot be said for their quality. The plethora of different ESG rankings has left investors frustrated and confused with their competing measurement methodologies. In fact, the latter vary so wildly that organisations have been able to cherry pick the most appealing providers (Murray, 2021; Li and Polychronopoulos, 2020). It also makes it difficult to compare one ESG ranking methodology with another (Berg et al., 2020). To illustrate, correlations between ESG ratings are on average 0.54 and range from 0.38 to 0.71. Sustainalytics and Vigeo Eiris have the highest level of agreement between each other, with a correlation of 0.71. The correlations of the environmental dimension are slightly lower than the overall correlations, with an average of 0.53 (Berg et al., 2020). This leads to capital markets to not adequately pricing the ultimate costs surrounding sustainable businesses. On top of this, the lack of a uniform measurement approach can cause reputational damage, financial loss and regulatory fines.

In general, the major rating providers follow similar processes for calculating their ratings. They use traditional research methodologies. [2]These include manual gathering of publicly available information and sending surveys to companies. Issuers also are given the opportunity to engage with rating providers to send feedback on the score given to them. Thus, ESG rankings appear to be manual, time-consuming, and prone to human bias and omissions.

The lack of consistency surrounding ESG scores can be explained by the multitude of choice in terms of disclosure standards and measurement methodologies. This paper takes as its premise the view that the limitations of the research methods used to generate ESG ratings are one of the biggest barriers to accuracy. In addition, current ESG score methodologies lack robust data, hence are failing to produce reliable ESG ratings.

Small and Wide Data. According to research by Gartner, 70% of organisations will shift their focus from big to small and wide data by 2025. [3] Small data refer to an approach that requires less data but still offers useful insights. By using "X analytics" techniques, wide data enable the analysis and synergy of a variety of small and varied, unstructured and structured data sources to enhance decision making. Small and wide data seek to solve challenges stemming from complex questions, scarce and disparate data, and historical data abruptly becoming obsolete and thus breaking AI models. To illustrate, breaking news articles can cause sudden change in sentiment surrounding an organisation, therefore making historical company data irrelevant.

Entity Salience. Salience is a linguistic term that refers to the importance or centrality of a word or phrase within a particular document. Entity salience scores are always relative to the analysed text. In general, scores closer to zero are less salient, while scores closer to one are highly salient (Wu et al., 2020).

B. Methodology

Category Collection: FTSE 100 Index. The Financial Times Stock Exchange 100 Index (FTSE 100 Index) is a share index tracking the 100 biggest companies by market capitalisation that are listed on the London Stock Exchange (LSE). [4] It is said to be the most used UK stock market indicator by investors. This paper takes the biggest eleven index constituents by market capitalisation as news search categories. [5] [6] The goal is to ensure the dataset is relevant for analysing the credentials of stock market companies. Since the focus of this paper is to produce a dataset with ESG relevance labels, the paper deems relevant news articles that include both a category and at least one ESG topic. For example, a news article about BP and an oil spill would be considered an environmental-related topic about the category "BP". BHP and miner union strike would be categorised as a social topic about the category "BHP". Royal Dutch Shell and lack of women in boardroom would refer to a governance topic with a category "Royal Dutch Shell".

ESG Topics. This paper considers as a topic any news article with ESG relevance. It can include any environmental, social, or governance-related news. Examples of ESG topics are provided in the Notion of ESG section.

Article Collection. News API is a REST API that returns JSON results for current and historic news articles. [7] This paper utilises it to retrieve news articles for each category. Being an established information retrieval method, it proved more suitable for the purposes of this research as it overcame the limitations associated with other news APIs and data collection techniques, such as RSS feed and web scraping. Namely, it allows to effortlessly collect historic articles and has a wide range of endpoints, including full content. It also facilitates reproducibility as the developer subscription is free. It ensures data robustness and non-bias as it returns results from over 80 000 news publications. In addition, it solves for data sparsity associated with ESG article collection.

The data collection process consists of two steps. The first one is collecting articles for each category and the second is dataset pre-processing. For consistency, the same endpoints,

[1] https://www.investopedia.com/terms/e/environmental-social-and-governance-esg-criteria.asp

[2] Based on the author's experience in capital markets and ESG and working with ESG rating providers.

[3] https://www.gartner.com/en/newsroom/press-releases/2021-05-19-gartner-says-70-percent-of-organizations-will-shift-their-focus-from-big-to-small-and-wide-data-by-2025

[4] https://www.investopedia.com/terms/f/footsie.asp

[5] As of 30 June 2021, https://www.londonstockexchange.com/indices/ftse-100/constituents/table

[6] The top 11 constituents were taken due to Royal Dutch Shell being listed twice.

[7] https://newsapi.org/

language, and time frame were used for each category. Category articles were extracted in a csv format. The downloads were performed in ten batches over a ten-day period due to subscription level restrictions. This produced ten csv files which were combined into one. News API has a historical data window of approximately three years. Thus, news articles in English were extracted within this time frame – between October 2018 and July 2021. To ensure data robustness, data were extracted by relevancy. A limit of 500 articles was set for each category. The following endpoints were extracted via News API for each article*: title, author, source, description, content (where available), publish date, and URL*. The following formatting changes were made to the data file to enhance understandability for future data use. Changes were made first in the Excel file and then using Python.

Duplicate news articles were removed using Python. After examining the new dataset, more duplicates were noticed. Different news publications reusing the same articles caused these news articles to be treated as different by Python. The duplicate rows were removed manually from the csv file.

The content of seven news articles had to be reduced. The aim was for each article to fit within a single Excel cell to avoid losing data during experiments. As a result, a limit of 4800 words per cell was applied in the dataset Excel file.

The "content" endpoint was renamed to "text". This was done in Python. A new column "Label" was also created for ESG labelling.

Articles were checked for personally identifiable information, particularly e-mail addresses. This was done by searching for symbols and domains commonly used in email addresses, i.e., "@" or ".com,". Author names were removed.

For the purposes of this study, only the "text", "Number" and "Label" columns were kept in the final dataset. All other columns were removed.

After iterating over each category, a total of 5000 articles was obtained. After removing duplicate articles, the final dataset consisted of 3913 articles. 1178 articles were labelled as ESG relevant, and 2735 articles were identified as irrelevant. Full article collection methodology is available in Appendix 1.

ESG Labelling. Article labelling was done manually. This method was adopted as it is considered the most precise for document annotations. A binary labelling method was proposed: 1 – articles related to ESG topics; 0 – articles not related to ESG topics.

Ethical Review. A developer News API license was obtained to download the news articles. Data were downloaded and used in accordance with News API Terms. [8] According to the provider, all data are publicly available. No personal data, such as user analytics or cookies were used in this study. News API is compliant with UK and EU data laws and directives. To illustrate, their privacy policy states that it "has been prepared in fulfilment of the obligations under Art. 10 of EC Directive n. 95/46/EC, and under the provisions of Directive 2002/58/EC, as revised by Directive 2009/136/EC, on the subject of Cookies." [9]. Thus, all data used for the purposes of this study is believed to be ethical.

C. The ESG-FTSE Dataset

The dataset described here consists of one readme file and a csv file. These files should be used in conjunction with each other to appropriately utilise the data. A summary and description of the files composing the dataset are available in Appendix 1.2. Descriptions of the variable names are provided in Appendix 1.3. Variables were named based on the research purpose. The values of the variables for ESG labels are represented as either a "0" for non-ESG or "1" for ESG relevant articles, as described in Methodology section E. The dataset is available in the "ESG-FTSE.zip" file provided with this paper. Basic dataset statistics are shown in Fig. 1.

TABLE I. ESG-FTSE DATASET STATISTICS

# raw articles	*5000*
# articles after pre-processing	3913
# articles with ESG relevance	1178
# articles with no ESG relevance	2735
# unique words	3892

Fig. 1. ESG-FTSE dataset statistics. Average length is shown in terms of words.

Technical Validation. Measures were taken to ensure the validity of the dataset, pre- and post-data collection. Sources of uncertainty and potential bias in the data are outlined below. Data quality and completeness checks were done before and after data collection. Manual spot checks were performed on the raw dataset file to ensure article content was available for all news articles. Next, missing value checks were also undertaken in Python. No blank content was discovered after the second pre-processing.

Usage Notes. There are six notebooks containing the Python code of dataset experiments, data extraction and data cleaning. They are available in standard ipynb format and may be imported into a variety of environments, including Google Colab and Jupyter. A list of all notebooks with descriptions is available in Appendix 1.2. ESG-FTSE is in a standard scv format. It is well-suited in its current form to be treated as factors or categories for research purposes.

IV. EXPERIMENT 1: AUTOMATIC TEXT CLASSIFICATION OF ESG RELEVANCE

A. Background

Automated Text Classification of Financial Texts. Traditionally, quantitative financial data have been essential to understanding an investment's sustainability potential. In recent years, alternative data, such as news articles, social media, image, and voice data, have become more important for assessing investment opportunities and financial market performance as they capture corporate information that falls outside the realm of traditional financial data (Hagenau et al., 2012; Kalyanaraman et al., 2014; Gupta et al., 2020; Shah et al., 2018; Shah et al., 2019). Studies have shown that such information, especially news articles, affects the value and performance of organisations (Shah et al., 2018; Shah et al., 2019). This in turn has boosted research in financial news analysis (Shah et al., 2019).

The volume of company-related data has seen an exponential growth over the past decade. It has been estimated

[8] https://newsapi.org/terms

[9] https://newsapi.org/privacy

that 80% of these data are unstructured, with text being one of the most common types of unstructured information. [10] The limitations of traditional research methods have made it difficult for investors to consider all available information. Traditional methodologies are labour-intensive, time-consuming and their objectivity and replicability are questionable. The advancement of AI computational techniques has made it possible to capture the value of vast amounts of text, such as news articles (Zhao et al., 2020; Luss and D'Aspremont, 2015). NLP is an AI approach that analyses text data. It has different sub methods. [11] This paper focuses on automated text classification applied to financial news. This is applied on news articles to capture ESG relevance.

Automatic text classification is a supervised learning method of taking an observation, extracting useful features, and classifying the observation into one of a discrete set of classes (Jurafsky and Martin, 2020). Classes can be binary or multi-class depending on the task. The algorithm comprises of text normalisation, feature extraction, dimensionality reduction, classifer selection, data training and testing, and evaluation. Text normalisation refers to a set of tasks that convert text to a more standard form. The most common techniques are tokenization, lemmatisation, stemming, sentence segmentation, and regular expressions. Feature extraction is commonly undertaken using techniques such as n-grams, term frequency and Word2Vec. Dimensionality reduction usually is performed using principal component analysis (PCA) and linear discriminant analysis. Another technique for performing nonlinear dimensionality reduction is t-distributed Stochastic Neighbor Embedding (t-SNE). [12]

Choosing a classifier is the most crucial step of the model. Choice of classifier depends on the nature of the task, thus there is no universal best classifier. Formally, there are two types of classifiers: generative and discriminative. Generative classifiers model how a class would generate input data. Given an observation, they predict the class most likely to have generated the observation. An example of such a classifier is Naïve Bayes Classifier (NBC) (Jurafsky and Martin, 2020). In contrast, discriminative classifiers like logistic regression (LR) learn what input features are most likely to discriminate between the different possible classes. Support Vector Machines (SVM) is another discriminative algorithm. Even though it requires more computational power compared to NBC and LR, it does not require much training data to produce good results. In fact, it is much faster and more accurate compared to NBC (ibid). SMV works is by drawing a hyperplane that divides the data into subspaces. These contain tags. The goal is to find the optimal hyperplane, i.e., the one with the largest distance between each tag. Due to its multi-dimensional nature, SMV works well with complex data (Kowsari et al., 2019). Discriminative models are more commonly used than generative ones due to their higher accuracy.

The next step of text classification is training and testing the model. The training set is used to fit the model and the test set is used to report on the algorithm's performance on unseen data. There are different ways of building a training and testing datasets (Jurafsky and Martin, 2020). Choosing the right one is largely dependent on the nature of the dataset. A common method is a fixed train/test split. The training set can be further broken down into a training and validation sets. The latter is used to obtain an early estimate of the model performance by allowing to tune hyperparameters. Another approach is to use a resampling method called cross-validation, i.e., the k-fold cross validation. The procedure has a single parameter called *k* referring to the number of groups that a given data sample is to be split into. Once the *k* parameter is set, the training and test datasets are built via random shuffling. Then a classifier is trained, and the error rate is computed on the test set. These steps are then repeated *k* number of times. The k-fold cross validation has several different variations. [13] An example is the stratified k-fold. Its aim is to ensure that each k-fold has the same proportion of observations with the class outcome value. Last, the evaluation step explains the performance of a classification model. Popular evaluation metrics are area under the ROC curve (AUC), recall, precision, F1 score, confusion matrix, decision function, and accuracy (Jurafsky and Martin, 2020). The latter appears to be the simplest method of evaluation.

Formally, automated text classification falls under three types of approaches: rule-based, machine learning and hybrid.[14] This paper explores the most common techniques for text classification of financial data that fall within these three categories. Rule-based models use manually crafted rules that detect semantically relevant features to identify relevant categories. Machine learning methodologies classify text based on past observations. Pre-labelled data are used to learn different text associations, i.e., a particular tag is expected for a particular piece of text. During training, the algorithm is fed with vectorised feature sets and tags to produce a classification model. During testing, the same approach is used with unseen data.

Both NBC and LR classifiers are common in sentiment analysis of financial data, social media, and stock price prediction.[14] They are computationally inexpensive, do not require big memory space or huge amount of data to produce good results. SVM is another popular algorithm for text classification. Like NBC and LR, it has been used in stock price prediction and sentiment analysis of financial texts (Gupta et al., 2020; Shah et al., 2019). Tree-based classifiers such as decision tree and random forest have also been used for the same classification tasks (Gupta et al., 2020). Deep learning (DL) algorithms have surpassed previous techniques in text classification. Subsequently, they are becoming increasingly popular for classifying financial data. DL is inspired by neural networks. There are two main types of DL architectures: Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). The biggest advantage of DL methods is their high accuracy and efficiency in terms of computational power. They also do not have a threshold for learning from training data, as opposed to SMV. However, despite some exceptions like a pre-trained BERT model, DL algorithms often require a large amount of training data, unlike traditional models. [15]

Related Work. Even though research in AI based ESG scores is still in its infancy, it has been gaining interest. In their journal paper, Aiba et al (2019) introduce a quantitative model that predicts a company's ESG ratings assigned by FTSE and

[10] https://www.capgemini.com/2018/08/reorganizing-unstructured-data/

[11] https://towardsdatascience.com/a-practitioners-guide-to-natural-language-processing-part-i-processing-understanding-text-9f4abfd13e72

[12] https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

[13] https://machinelearningmastery.com/k-fold-cross-validation/

[14] https://monkeylearn.com/text-classification/

[15] https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270

MSCI. It deploys image recognition and NLP techniques to assess how well the FTSE and MSCI scoring methodologies work and the impact of disclosure items on individual ESG score providers.

In another paper, the authors focus on detecting historical ESG trends in corporate earning transcripts (Raman et al., 2020). Another study leverages deep learning and ESG ratings into a portfolio optimisation model (Sokolov et al., 2021). In her thesis paper, Napier (2019) explores the use of machine learning methods on CSR reports. In a different study, Shahi et al (2014) apply different classifiers to predict sustainability scores. In their paper, Hisano et al (2020) discuss adverse media coverage to predict companies likely to be blacklisted in sustainable investment practices. Khan (2019) proves that ESG metrics can determine stock returns in a global investment universe. Nematzadeh et al (2019) propose real-time event detection techniques and sentiment analysis are combined to identify Twitter controversies around companies and associate these with stock performance. Thomson Reuters data are used by Ribando and Bonne (2010) to illustrate that socially responsible companies create long-term shareholder value. The predictive power of CSR information on financial returns has been explored by Ghoul et al (2011). Krueger et.al. (2020) conclude that climate change has huge financial implications to investors. The study by Guo et al (2020) explores the effect of ESG financial news on market volatility. Keyword-based analysis of ESG characteristics has also attracted research (Brown, 2015). However, the current research argues that such approaches are prone to error as they do not capture semantic understanding of the text. More advanced methods, such as DL, have improved semantic research (Pennington et al., 2014; Goel et al., 2020; Vaswani et al., 2017; Peters et al., 2018).

Summary. While analysing company narrative and social media text carry relevant ESG information, this paper takes the view that such approaches have limitations in terms of data robustness and objectivity. To illustrate, detecting ESG relevance from corporate disclosure and earning call reports hampers ESG score objectivity and accuracy by excluding other sources of important information, such as news articles. In addition, not all companies produce Corporate Social Responsibility (CSR) reports or include such sections in their annual reports. Companies tend not to voluntarily disclose negative sentiment about themselves either. Another limitation of the relevant studies is the lack of generalisability of their models to out-of-corpora models. This paper aims to address these limitations by creating an ESG news article dataset that is objective and applicable to other corpora domains. This in turn will encourage further research in ESG investing.

B. Methodology

Supervised Learning: Text Classification. Three classification experiments were undertaken on ESG-FTSE. Experiment 1 used a stratified cross-validation approach with text pre-processing. Experiment 2 adopted a training/test split and n-grams. Experiment 3 was an extension of Experiment 1, with n-grams added to the pre-processing step. The same classifier was utilised in all three experiments. Relevant notebook and full list of Python packages used for the experiments is presented in Appendix 1.2 and 2.1.

To facilitate text classification, the corpus was first pre-processed in Experiment 1. The NLTK Python package was used to change the text to lower case, tokenize it, remove stopwords and unwanted characters, such as numbers and special characters, then lemmatize and stem it. Next, data were vectorised via the TF-IDF approach. This created unigrams.

A 5-fold stratified cross-validation was applied to the model. Due to the imbalanced nature of the dataset classes, it was important to ensure that the data were split randomly while maintaining the same class distribution in each subset. The *k fold* number was chosen based on the class ratio. To illustrate, the class ratio of the dataset was 0.30. Splitting the dataset into five folds ensured that the fold values were similar to the class ratio (Appendix 2.2). In general, SVM classifiers are produce highly accurate results for binary classification problems. Thus, a SVM classifier was chosen to test the algorithm. The following SVM parameters were used: the regularization parameter *C* was kept default; a linear *kernel* was applied due to the binary nature of the task; the *degree* of the polynomial kernel function was set to its default value; *gamma* was set to “auto”. At testing, data were divided with a 67:33 split ratio. Last, evaluation was undertaken using ROC curve and ROC AU score, accuracy, precision, recall, F1 score measures, decision function and confusion matrix.

Experiment 2 was conducted to determine if a different data split, n-grams and limited pre-processing would improve classification accuracy. The corpus was split into a training and validation datasets. CountVectorizer was used to count vectors as features and transform the training and validation datasets into vectorizer objects. Next, TF-IDF was used for feature extraction at different levels: word, n-grams (from bi- to four-grams), and character. Last, the model was built with a SVM classifier. The SVM model parameters from the first experiment were used. Accuracy, precision, and recall were printed.

The code from Experiment 1 was replicated in Experiment 3. In addition, a different tokenizer (RegexpTokenizer) was added, as well as bi-grams, trigrams and four-grams.

C. Results

Basic experiment evaluation metrics are presented in Fig. 2. Across all experiments, Experiment 1 outperformed the other experiments across almost all evaluation metrics. Experiment 2 produced the worst results. To illustrate, its recall value indicates a big number of missed positive predictions. In terms of accuracy, Experiment 1 was 3.5% higher than Experiment 2 but only marginally better compared to Experiment 3 (88.6% vs. 87.8% respectively). Inspection of error rates of Experiment 1 and Experiment 3 revealed that the former was both more specific, i.e., how selective the classifier is in predicting positive values, and sensitive, i.e., how often it detects true positives. In addition, it had a lower false positive rate. To illustrate, the confusion matrices were analysed for both classification experiments. In addition, their specificity, false positive rate, and decision functions were

TABLE II. CLASSIFICATION RESULTS

Exp. #	**Evaluation Metrics in %**			
	Accuracy	***Precision***	***Recall***	***F1***
Exp.1	88.6	86.3	73.0	79.1
Exp.2	85.1	82.4	65.4	73.0
Exp.3	87.8	83.3	74.0	78.4

Fig. 2. Text classification results. Results are rounded to 1 decimal place.

computed. Evaluation metrics are presented in Appendix 2.3. Their comparison demonstrated that the predicted sample of Experiment 1 lies closer to the decision boundary. Its specificity was higher while its false positive rate was lower. Analysis of the ROC curves and AUC values showed very high similar outputs: 0.93 for Experiment 1 vs. 0.94 for Experiment 3 (Appendix 2.3). In addition, Experiment1 has a higher F1 and Precision scores compared to Experiment3., despite a slightly lower recall value. Thus, it proved to be the best performing model.

D. Discussion

The importance of feature selection and dataset characteristics in text classification was confirmed by the experiments. The experiments also demonstrated that ESG-has small data characteristics. Despite that, it was shown that insightful information can be obtained from ESG-FTSE.

The ability of the model in Experiment 1 to yield better results is likely attributable to its model structure and use of unigrams. ESG-FTSE has an imbalanced classification problem, thus requiring a carefully crafted classifier model to avoid poor predictive performance. To test this notion, a unigram model with a stratified 5-fold split (Experiment 1) was compared with a standard training/validation split model with uni-grams (Experiment 2), and an identical to Experiment 1 SVM model but with added n-grams. In general, n-grams tends to improve accuracy. Interestingly, neither of the n-gram algorithms outperformed the model in Experiment 1. To get a better understanding of performance, closer error analysis was undertaken. As described in the Results section, Experiment 1 was less prone to error. This suggests that the n-gram models might be overfitting. For Experiment 2, it could possibly be attributed to the training/validation split as it designed around the assumption of an equal number of examples for each class. This typically results in inaccurate classification, especially of the minority class. This poses a challenge because, just like other imbalance classification problems, the minority class is more important for the current study. An alternative explanation is n-grams causing overfitting in both Experiment 2 and Experiment 3 as n-grams tend to be less generalisable. To illustrate, the bi-gram “shareholder revolt” would be more generalisable compared to the trigram “by shareholder revolt”.

V. Experiment 2: Topic modelling of the ESG Corpus

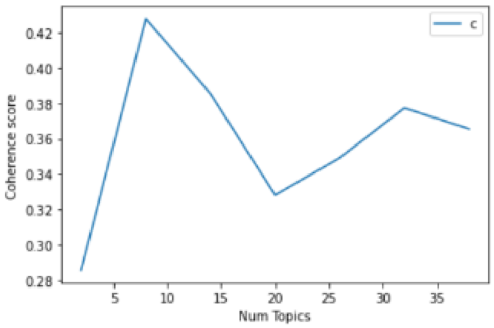
A. Background

Topic Modelling in Semantic Networks. Topic modelling is an unsupervised probabilistic algorithm that considers the problem of modelling discrete data, such as text corpora. The goal is to discover the main topics that occur in a set of documents by reducing their dimentionality. One of the primary applications of topic modelling is understanding customer opinions, complaints, reviews, feedback, news stories. The topics are represented by clusters of words (Blei et al., 2003). Originally, topic modelling was developed as a text-mining tool. The popular TF-IDF scheme is an example of an early topic model. Information retrieval (IR) researchers have also proposed a basic topic model methodology used in Internet search engines. As discussed by Blei et al (2003), other more advanced topic models, known as mixture models, are latent semantic indexing (LSI) and probabilistic LSI (pLSI). While these models are useful, they do not capture the exchangeability of both documents and words. One such model proposed by Blei et al (2003) is the latent Dirichlet allocation (LDA) model. It is generative latent variable model that represents documents as random mixtures over latent topics, where each topic is characterised by a distribution over words. It creates both a topic per document model and words per topic model. They are modelled as Dirichlet distributions. Each document is modelled by a multinomial distribution over topics, and each topic is modelled by a multinomial distribution over words. LDA assumed that documents are produced by a mixture of topics and returns words based on their probability distribution. It also assumes that words are interconnected in corpora, i.e., they are indicative of a topic. Thus, choosing the right corpus for a research task is crucial. There are several extensions to LDA. [16] Due to its unsupervised training process, topic models prove challenging to evaluate. In general, there are five approaches commonly used for evaluation: eye-balling models, human judgement, extrinsic metrics, and intrinsic metrics. [17] Common intrinsic metrics are perplexity and coherence.

B. Methodology

Topic Modelling. Topic modelling was performed by building a LDA model using a Gensim and mallet implementation. The code was adapted from a tutorial. [18] A relevant notebook and Python packages are presented in Appendix 1.2 and 3.16. All experiments were evaluated using an intrinsic metric (coherence score C_v), t-SNE, entity salience, and topic representation techniques. The NLTK and Spacy packages were utilised for data cleaning and tokenization. Next, bigrams and trigrams were created and lemmatized. To build the LDA base model, a dictionary and a corpus were created. The dictionary consisted of the lemmatized data. The corpus was the mapping of the unique ID for each word in the document and its frequency. To improve the accuracy of the default LDA model and allow reproducibility, a series of sensitivity tests were performed. To illustrate, additional pre-processing was done by tuning the *filter_extremes* hyperparameter to remove the most common and the rarest words in the corpus. Additionally, tests were performed to determine the most optimal number of topics.

Fig. 3.Coherence score graph for k number of tests. Choosing the optimal number of topics via sensitivity tests.



[16] https://www.analyticssteps.com/blogs/introduction-linear-discriminant-analysis-supervised-learning

[17] https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[18] https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

TABLE III. LDA MODEL SENSITIVITY TESTS

Test #	Hyperparameters				
	Alpha	*Beta*	*filter_extremes*	*Perplexity*	*Coherence c_v*
1	0.01	0.9	no_below=10, no_above=0.20	-7.597	0.538
2	0.01	0.5	no_below=10, no_above=0.15	- 7.670	0.565
3	0.03	0.5	no_below=10, no_above=0.15	-7.652	0.597
4	0.05	0.5	no_below=10, no_above=0.15	-7.673	0.552
5	0.04	0.5	no_below=10, no_above=0.15	- 7.665	0.572
6	0.04	0.4	no_below=10, no_above=0.15	- 7.668	0.551

Fig. 4. LDA model sensitivity tests with k=20. Perplexity and coherence scores are rounded to three decimal places.

Number of topics (*k*) is one of the most important LDA model inputs. Extracting the right number of topics largely depends on the dataset characteristics. [19] [20]. There are several ways of determining the most optimal number of topics.[18] For the purposes of this study, the *k* value was selected by building seven LDA models with different *k* values and comparing their coherence values (Fig. 3, Appendix 3.1). A limit of forty LDA models was set. The optimal number of topics was chosen based on the highest coherence value achieved on the final LDA model. Results are presented in Appendix 3.1. Thus, based on the test results, *k=32* was initially tested on the main LDA model because it marked the end of a rapid growth of the coherence score, i.e., the coherence value was increasing with the number of topics. It achieved a coherence score of 0.52. A second sensitivity test was run with the highest coherence score obtained during sensitivity testing, i.e., *k=14*. It produced a coherence value of 0.57. Last, a test with *k=20* was performed. It had the highest coherence score of 0.60 after tuning the hyperparameters described below. Thus, *k=20* was used in the final LDA model. A perplexity score was also computed for completeness, although it was not used in evaluation. In general, the lower the perplexity value, the better.

In addition, sensitivity tests were performed on the following hyperparameters to improve accuracy: *random_state, update_every, chunksize, passes, alpha, eta and per_word topics*. The most optimal hyperparameter values were chosen based on the highest coherence score achieved on the LDA model (Fig. 3., Test 3). All sensitivity test results are presented in Fig.3. Last, the model was run with the selected parameter values. It achieved a coherence score of 0.60. The topics were visualised via t-SNE (Fig.6), matplotlib, and pyLDAvis package's interactive chart (Fig.5.). Each bubble on the plot represents a topic. The larger the bubble, the more prevalent the topic. Additionally, a bar chart representing the top 30 most salient keywords that form a selected topic is presented in Appendix 3.2.

Another set of experiments were performed on the final LDA. First, the dominant topic in each sentence was determined by finding the topic number with the highest percentage contribution in the corresponding news article. Second, the most representative news article for each topic was identified. Last, the topic distribution across news article was computed. Notebook reference and illustrated results are available in Appendix 1.2 and 3.3-3.5.

Several experiments were conducted for the top four topics only. The code was adopted from a tutorial. [21] Changes to the final LDA model parameters were made due to the smaller number of topics. The altered model parameters are available in the relevant notebook (Appendix 1.2). The following illustrations and outputs are shown in Appendix 3.7-3.15: dominant topic and its percentage contribution in each document; the most representative sentence for each topic; frequency distribution of word counts in documents; word clouds of top n keywords in each topic; word count and

Fig. 5. LDA model Intertopic Distance Map.

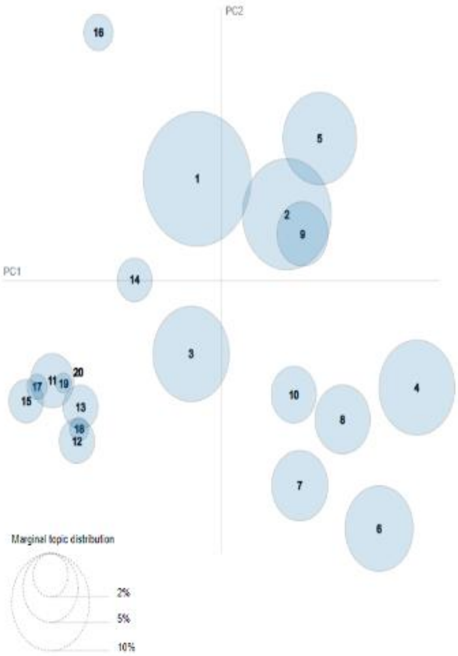
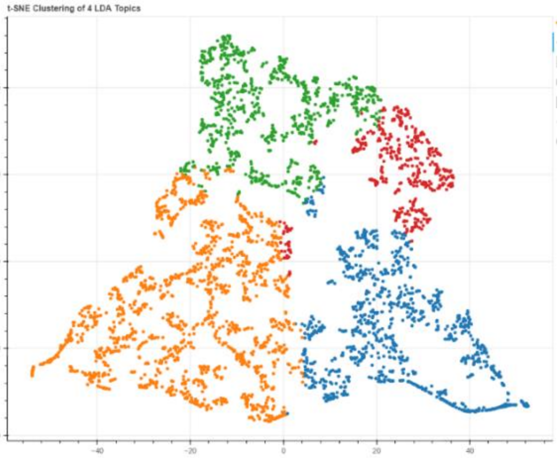


Fig. 6. Top Four Topics: t-SNE clustering chart (below)



[19] https://towardsdatascience.com/unsupervised-nlp-topic-models-as-a-supervised-learning-input-cf8ee9e5cf28

[20] https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[21] https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/

importance of topic keywords; sentence chart coloured by topic; the most discussed topics in documents; t-SNE clustering chart (Fig. 6); a second pyLDAVis interactive chart.

Topic Modelling for Classification. Topic modelling can also be used as a supervised learning method. For example, document classification. In this experiment LDA was used to classify sample and unseen articles. The code was adopted from an online tutorial.[22] Data were pre-processed using the Gensim and NLTK libraries. Two LDA models were created. LDA Bag-of-words model was computed by creating a bag-of-words dictionary containing word count and frequency. Next, a second LDA model was set up using TF-IDF to transform the corpus. Performance evaluation was run by classifying sample news articles via each LDA model. Last, LDA Bag-of-words model was tested on unseen data. Outputs are shown in Appendix 3.16. The notebook is referenced in Appendix 1.2.

C. *Results*

Experiment outputs are presented in Appendix 3. Results demonstrated that ESG-FTSE can be successfully utilised in both unsupervised and supervised topic modelling tasks. In addition, the model overcame the small data problem associated with the dataset by managing to extract meaningful insights about FTSE 100 constituents with ESG relevance. Last, the topic modelling experiments provided a thorough context and description of the dataset.

To illustrate, the topic model achieved a coherence score of 0.60. Given the dataset characteristics and number of topics, this paper takes the view that it is a high coherence score. In addition, all topic modelling experiments produced some ESG-related results. For example, the most representative document for topic 2 clearly contains environmental keywords: *climate, bank, carbon, emission, cigarette* (Appendix 3.4). Useful statistics were also extracted for topics. To illustrate, the most salient words for each topic are shown in Appendix 3.2 followed by the most dominant topic in each document and sentence, topic distribution across documents, word count. In addition, the LDA model correctly classified an unseen news article, achieving an 89% score (Appendix 3.15).

A closer look of the four biggest topics provided a more detailed information about data density and volume (Appendix 3.6-3.14). For example, the mean of the distribution of document word counts was 125 (Appendix 3.14). Word cloud visualisations of the most dominant topics revealed that one of them is ESG-related (Appendix 3.9). The experiment shown in Appendix 3.12 showed that one of the most dominant document topics was about plastic. This suggests ESG relevance.

D. *Discussion*

The LDA model and classification were evaluated using multiple techniques. As discussed in the Results section, the experiments indicated a robust and objective LDA model that can classify unseen data.

Consistent labelling, evaluation metrics, clearly defined topics, and parameter and hyperparameter tuning were pivotal in building a high performing LDA model. Finding the most optimal values required implicit knowledge not only of topic modelling techniques but also of the dataset characteristics and the task goals. To illustrate, probabilistic topic models like LDA always produce topic outputs. However, making them useful and meaningful for this research demanded capturing the correct information, i.e., the minority ESG class. Despite the class imbalance of the dataset, both the topic modelling and the topic model classification experiments were successful in extracting relevant ESG information. Sensitivity tests ensured output relevance and robustness as measured by the coherence score.

Visualising the topics also indicated a well-built model. Fig.4. shows that most topic bubbles are well defined, of good size and scattered around the plot. In general, the larger and more scattered the bubbles and the higher the distance between them, the better. In addition, the t-SNE plot of the four most dominant topics indicates that the most similar documents are grouped together in well-defined clusters (Fig.6). The robustness of the model was also demonstrated by the most salient words for each topic (Appendix 3.1). It proved that the model could capture ESG data. Another visualised experiment determined that some of the most important words by weight and word count were ESG keywords, such as *emission* and *plastic* (Appendix 3.10).

VI. Conclusion

The greatest contribution of this study is presenting a free reproducible dataset to facilitate the sharing of data in a standardised framework. Being the first dataset with ESG relevance, ESG-FTSE provides a novel solution to the bias associated with ESG evaluation metrics. This study believes that the dataset could potentially boost research in the ESG subdomain. A crucial extension of this work will be expanding the dataset to include all FTSE 100 constituents, more news publication sources and other data types. Another possible avenue for further work would be the creation of a semi-automated ESG scoring system. The ESG-FTSE dataset was described in this paper. Experiments validated the proposed hypothesis that the dataset would be useful for ESG research purposes. To illustrate, the current study has shown that it is possible to make accurate ESG classifications (88% accuracy, 79% F1 score) and derive useful ESG topic information about FTSE constituents. It should be noted that due to the limited available ESG news about the categories, the data are representative only of the FTSE companies as they are described by the news publications obtained from News AI. Another limitation is that the dataset only includes news sources in English. A possible avenue to explore is including news articles in other languages.

Class imbalance due to data sparsity was revealed by the present study. Instead of trimming the dataset or boosting the minority class, this paper trained on all possible instances to maximise coverage. The evidence presented confirmed that small data can be insightful in obtaining relevant ESG insights.

References

Aiba, Y., Ito, T. and Ibe., Y. (2019) Network structure in ESG ratings suggests new corporate estrategies: evolving AI technology to quantify qualitative data, *Securities Analysis Journal* [online]. Available from: https://www.saa.or.jp/english/publications/2020_aiba_ito_ibe.pdf [Accessed July 2021].

[22] https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24

Berg, F., Kölbel, J., and Rigobon, R. (2020) Aggregate Confusion: The Divergence of ESG Ratings, SSRN [online]. Available from: https://ssrn.com/abstract=3438533 [Accessed 13 July 2021].

Blei, D., Edu, B., Ng, A., Jordan, M. and Edu, J. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, pp.993–1022. [online]. Available from: https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf. [Accessed 22 June 2021].

Brown, M. (2015) Managing Nature–Business as Usual: Resource Extraction Companies and Their Representations of Natural Landscapes, *Sustainability*, 7(12), pp.15900–15922.

Ghoul, S., Guedhami, O., Kwok, C. and Mishra, D. (2011) Does corporate social responsibility affect the cost of capital? *Journal of Banking & Finance*, 35(9), pp.2388–2406. [online] Available from https://econpapers.repec.org/article/eeejbfina/v_3a35_3ay_3a2011_3ai_3a9_3ap_3a2388-2406.htm. [Accessed 20 July 2021].

Guo, T., Jamet, N., Betrix, V., Piquet, A. and Hauptmann, E. (2020) ESG2Risk: A Deep Learning Framework from ESG News to Stock Volatility Prediction. arXiv:2005.02527 [cs, q-fin]. [online] Available from https://arxiv.org/abs/2005.02527 [Accessed 18 June 2021].

Hagenau, M., Liebmann, M., Hedwig, M. and Neumann, D. (2012) Automated News Reading: Stock Price Prediction Based on Financial News Using Context-Specific Features, 45th Hawaii International Conference on System Sciences, pp. 1040-1049, doi: 10.1109/HICSS.2012.129.

Hisano, R., Sornette, D. and Mizuno, T. (2020) Prediction of ESG compliance using a heterogeneous information network. *Journal of Big Data*, 7(1). [online]. Available from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00295-9 [Accessed 22 July 2021].

Huang, K. and Li, Z. (2011) A multilabel text classification algorithm for labeling risk factors in SEC form 10-K, *ACM Trans. Manage. Inf. Syst*. vol. 2, no. 3, Article 18, pp. 19. doi: https://doi org.ezproxy.library.qmul.ac.uk/10.1145/2019618.2019624.

Jurafsky, D. and Martin, J. (2020) *Speech and Language Processing An Introduction to Natural Language Processing*, Computational Linguistics, and Speech Recognition Third Edition draft. [online] . Available at: https://web.stanford.edu/~jurafsky/slp3/ed3book_dec302020.pdf.

Kalyanaraman, V., Kazi, S., Tondulkar, R. and S. Oswal. (2014) Sentiment Analysis on News Articles for Stocks, 8th Asia Modelling Symposium, pp. 10-15, doi: 10.1109/AMS.2014.14.

Khan, M. (2019) Corporate Governance, ESG, and Stock Returns around the World, *Financial Analysts Journal*, 75(4), pp.103–123.

Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L. and Brown, D. (2019) Text Classification Algorithms: A Survey. *Information*, 10(4), p.150. [online] Available from: https://arxiv.org/pdf/1904.08067.pdf. [Accessed 20 June 2021].

Li, F. and Polychronopoulos, A. (2020) What a Difference an ESG Ratings Provider Makes!, Research Affiliates [online]. Available from https://www.researchaffiliates.com/en_us/publications/articles/what-a-difference-an-esg-ratings-provider-makes.html [Accessed 20 June 2021].

Luss,R. and D'Aspremont, A. (2015) Predicting abnormal returns from news using text classification, Quantitative Finance, vol. 15, no. 6, pp. 999–1012. doi: 10.1080/14697688.2012.672762.

Mooney, A. (2021) Investment industry at 'tipping point' as $43tn in funds commit to net zero, Financial Times [online]. Available from https://www.ft.com/content/e943869b-7afd-4eea-8e0c-6ba3991bc5e3 [Accessed 6 July 2021].

Murray, S. (2021) Measuring what matters: the scramble to set standards for sustainable business, Finanl Times [online]. Available from https://www.ft.com/content/92915630-c110-4364-86ee-0f6f018cba90 [Accessed 3 July 2021].

Napier, E. (2019) Technology Enabled Social Responsibility Projects and an Empirical Test of CSR's Impact on Firm Performance. Ph.D. Thesis, Georgia State University, Atlanta, GA, USA [online]. Available from https://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1052&context=marketing_diss. [Accessed 15 July 2021].

Nematzadeh, A., Bang, G., Liu, X. and Ma, Z. (2019) Empirical Study on Detecting Controversy in Social Media. arXiv:1909.01093 [cs, stat]. [online] Available at: https://arxiv.org/abs/1909.01093 [Accessed 1 August 2021]. 2019, arXiv:1909.01093.

Ribando, M. and Bonne (2010) New Quality Factor: Finding Alpha With ASSET4 ESG Data; Starmine Research [online]. Available from https://www.thomsonreuters.com/content/dam/openweb/documents/pdf/tr-com-financial/report/starmine-quant-research-note-on-asset4-data.pdf [Accessed 11 May 2021].

Note; Thomson Reuters: New York, NY, USA, 2010; Volume 31.

Shah, D., Isah, H. and Zulkernine, F. (2019) Stock Market Analysis: A Review and Taxonomy of Prediction Techniques, International Journal of Financial Studies, vol. 7, no. 26. doi: https://doi.org/10.3390/ijfs7020026.

Shah, D., Isah, H. and Zulkernine, F. (2018) Predicting the Effects of News Sentiments on the Stock Market, 2018 IEEE International Conference on Big Data, Big Data, pp. 4705-4708, doi: 10.1109/BigData.2018.8621884.

Shahi, A.M., Issac, B. and Modapothala, J.R. (2014) Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring. *International Journal of Computational Intelligence and Applications*, 13(01), p.1450006 [online. Available from https://www.worldscientific.com/doi/abs/10.1142/S1469026814500060 [Accessed 16 July 2021].

Sokolov, A., Mostovoy, J., Ding, J. and Seco, L. (2021) Building Machine Learning Systems for Automated ESG Scoring, The Journal of Impact and ESG Investing, vol. 1, no. 3, pp. 39-50. doi: https://doi.org/10.3905/jesg.2021.1.010.

Vo, N., He, X., Liu, S. and Guandong, X. (2019) Deep Learning for Decision Making and the Optimization of Socially Responsible Investments and Portfolio, Decision Support Systems, vol. 124. Doi: 113097. 10.1016/j.dss.2019.113097.

Wu, C., Kanoulas, E. and de Rijke, M. (2020) WN-Salience: A Corpus of News Articles with Entity Salience Annotations. [online]Available at: https://aclanthology.org/2020.lrec-1.257 [Accessed 30 May 2021].

Zhao, W., Zhang, G., Yuan, G., Liu, J., Shan H. and Zhang, S. (2020) The Study on the Text Classification for Financial News Based on Partial Information, IEEE Access, vol. 8, pp. 100426-100437, 2020, doi: 10.11.

Appendix

Appendix 1 – ESG-FTSE article collection

1.1 Article Collection Methodology

Category search terms	ASTRAZENECA, UNILEVER, DIAGEO, HSBC, GLAXOSMITHKLINE, RIO TINTO, BP, BRITISH AMERICAN TOBACCO, ROYAL DUTCH SHELL, BHP
Time period	30/10/2018 – 31/07/2021
ESG labels	1=ESG-relevant, 0=Non-ESG
News Article Source	News API
Raw Endpoints	title, author, source, description, content (where available), publish date, and URL
Final Endpoints	content
Added columns	content was renamed to text, a Number column containing article number was added.
Excel file spot check	Spot checks to ensure content is available and there are no blanks.
Python: article duplicate removal	Duplicate articles were removed using Python
Excel: article duplicate removal	Manual duplicate removal by sorting the
Word limit	Word limit of 4800 words was set to seven articles in Excel.
Personally identifiable information	Personally identifiable information was removed via Python.

1.2 Summary of ESG-FTSE files and Python notebooks with descriptions

ESG-FTSE dataset	csv file containing the dataset
README file	Txt file containing instructions
News API_extraction.ipynb	Article extraction
cleaningup_text.ipynb	Data clean-up
SVM_exp_stratified_kfold_ngrams_3_experiments.ipynb	Text classification
Topic_modelling.ipynb	Topic modelling and experiments
LDA_with_tfidf.ipynb	Topic modelling for classification and experiments
topic_visualisation.ipynb	Visualisation of top four topics

1.3 Description of ESG-FTSE variable names

Number	Article ID
Text	Article content
Label	ESG Label. 1=ESG relevant; 0=non-ESG

Appendix 2 – Text Classification

2.1 Experiment 1, 2 and 3: Python packages

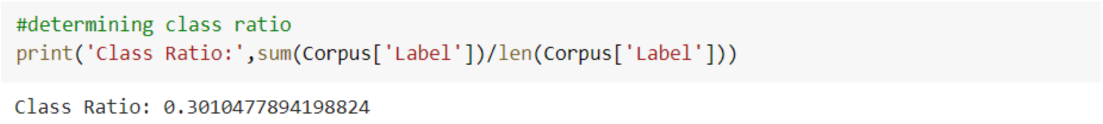
Commented code is available in notebook “SVM_exp_stratified_kfold_ngrams_3 experiments.ipynb”. The table below presents packages used in all three experiments.

Main packages	pandas, unicodedata, nltk, os, csv, wordnet, import numpy, defaultdict, shuffle, decomposition, ensemble, Pipeline
Pre-processing packages	NLTK, sklearn, preprocessing from sklearn
Lowercase	NLTK

Stopwords and unwanted character removal	word_tokenize from NLTK, RegexpTokenizer
	nltk.download('punkt')
	nltk.download('wordnet')
	nltk.download('stopwords')
	nltk.download('averaged_perceptron_tagger')

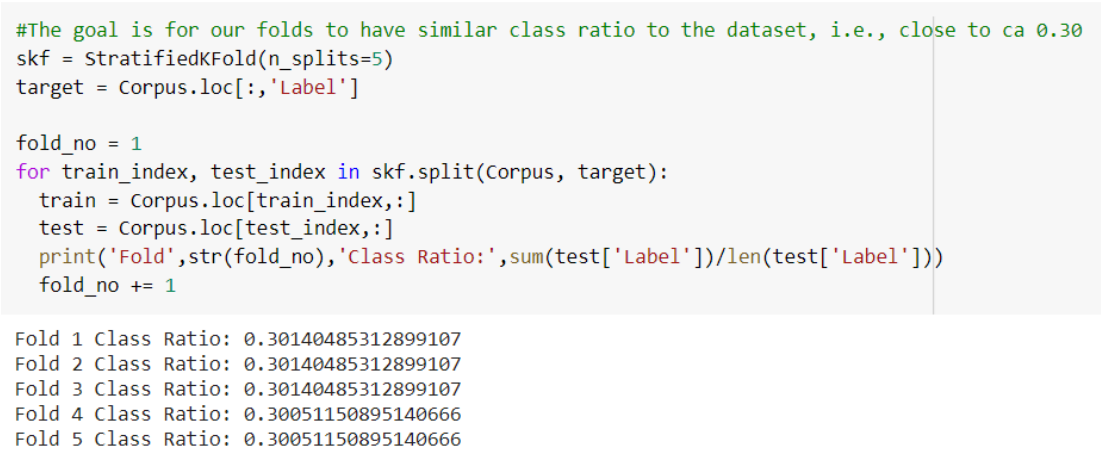
Pos tags	pos_tag from NLTK
Tokenization	word_tokenize
Lemmatization	WordNetLemmatizer
Stemming	WordNetLemmatizer
Feature extraction	TfidfVectorizer, CountVectorizer
n-grams	CountVectorizer
Model building	model_selection ,linear_model,
Data splitting	train_test_split,cross_val_score
Stratified k-fold	StratifiedKFold from sklearn.model_selection
k-fold selection	Based on class ratio
Classifier	svm from sklean
Evaluation metrics	From sklearn: metrics, precision_recall_fscore_support, plot_roc_curve
Illustration	Matplotlib, pyplot

2.2 Experiment 1 and 2: code snippet of determining k-fold number



```
#determining class ratio
print('Class Ratio:',sum(Corpus['Label'])/len(Corpus['Label']))
```

```
Class Ratio: 0.3010477894198824
```



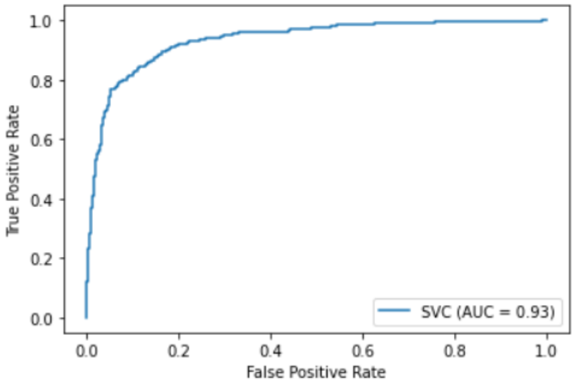
```
#The goal is for our folds to have similar class ratio to the dataset, i.e., close to ca 0.30
skf = StratifiedKFold(n_splits=5)
target = Corpus.loc[:,'Label']

fold_no = 1
for train_index, test_index in skf.split(Corpus, target):
  train = Corpus.loc[train_index,:]
  test = Corpus.loc[test_index,:]
  print('Fold',str(fold_no),'Class Ratio:',sum(test['Label'])/len(test['Label']))
  fold_no += 1
```

```
Fold 1 Class Ratio: 0.30140485312899107
Fold 2 Class Ratio: 0.30140485312899107
Fold 3 Class Ratio: 0.30140485312899107
Fold 4 Class Ratio: 0.30051150895140666
Fold 5 Class Ratio: 0.30051150895140666
```

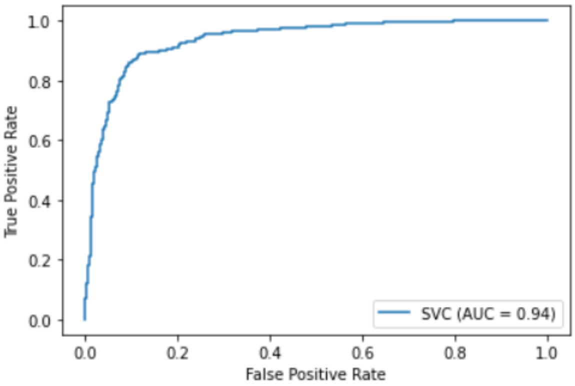
2.3 Experiment 1 and 3: evaluation metrics

Experiment 1and 3: ROC curves

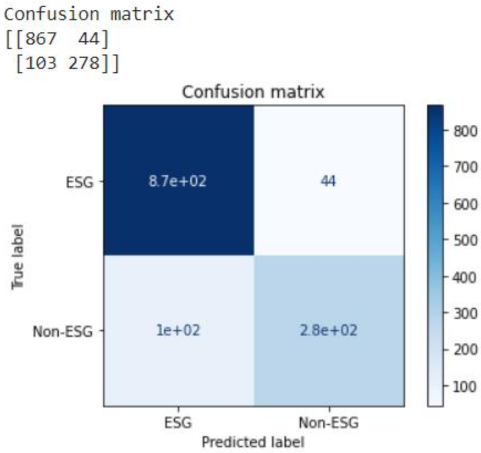
Experiment 1 ROC curve



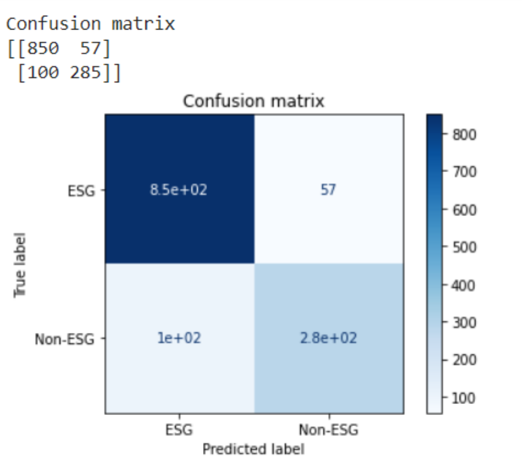
Experiment 3 ROC curve



Experiment 1and 3: Confusion matrices



```
Confusion matrix
[[867  44]
 [103 278]]
```



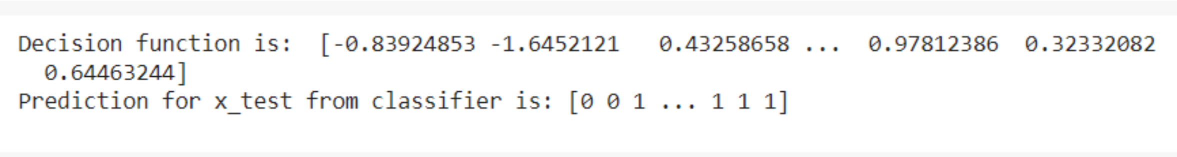
```
Confusion matrix
[[850  57]
 [100 285]]
```

Experiment 1 confusion matrix

Experiment 3 confusion matrix

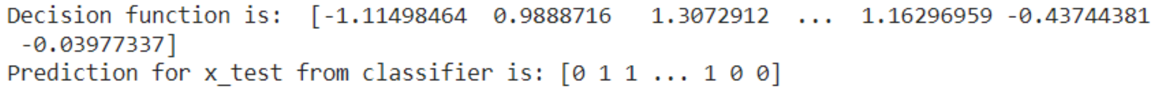
Experiment 1 and 3: Decision functions

Experiment 1 decision function:



```
Decision function is:  [-0.83924853 -1.6452121   0.43258658 ...  0.97812386  0.32332082
  0.64463244]
Prediction for x_test from classifier is: [0 0 1 ... 1 1 1]
```

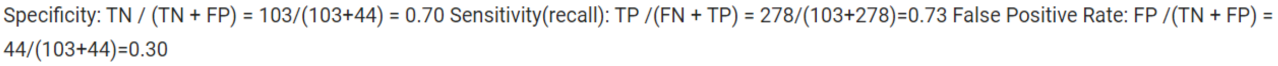
Experiment 3 decision function



```
Decision function is:  [-1.11498464  0.9888716   1.3072912  ...  1.16296959 -0.43744381
 -0.03977337]
Prediction for x_test from classifier is: [0 1 1 ... 1 0 0]
```

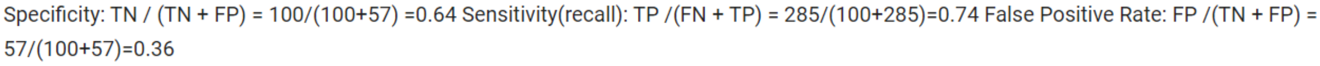
Experiment 1 and 3: specificity, recall, false positive rate

Experiment 1:



Specificity: TN / (TN + FP) = 103/(103+44) = 0.70 Sensitivity(recall): TP /(FN + TP) = 278/(103+278)=0.73 False Positive Rate: FP /(TN + FP) = 44/(103+44)=0.30

Experiment 3:



Specificity: TN / (TN + FP) = 100/(100+57) =0.64 Sensitivity(recall): TP /(FN + TP) = 285/(100+285)=0.74 False Positive Rate: FP /(TN + FP) = 57/(100+57)=0.36

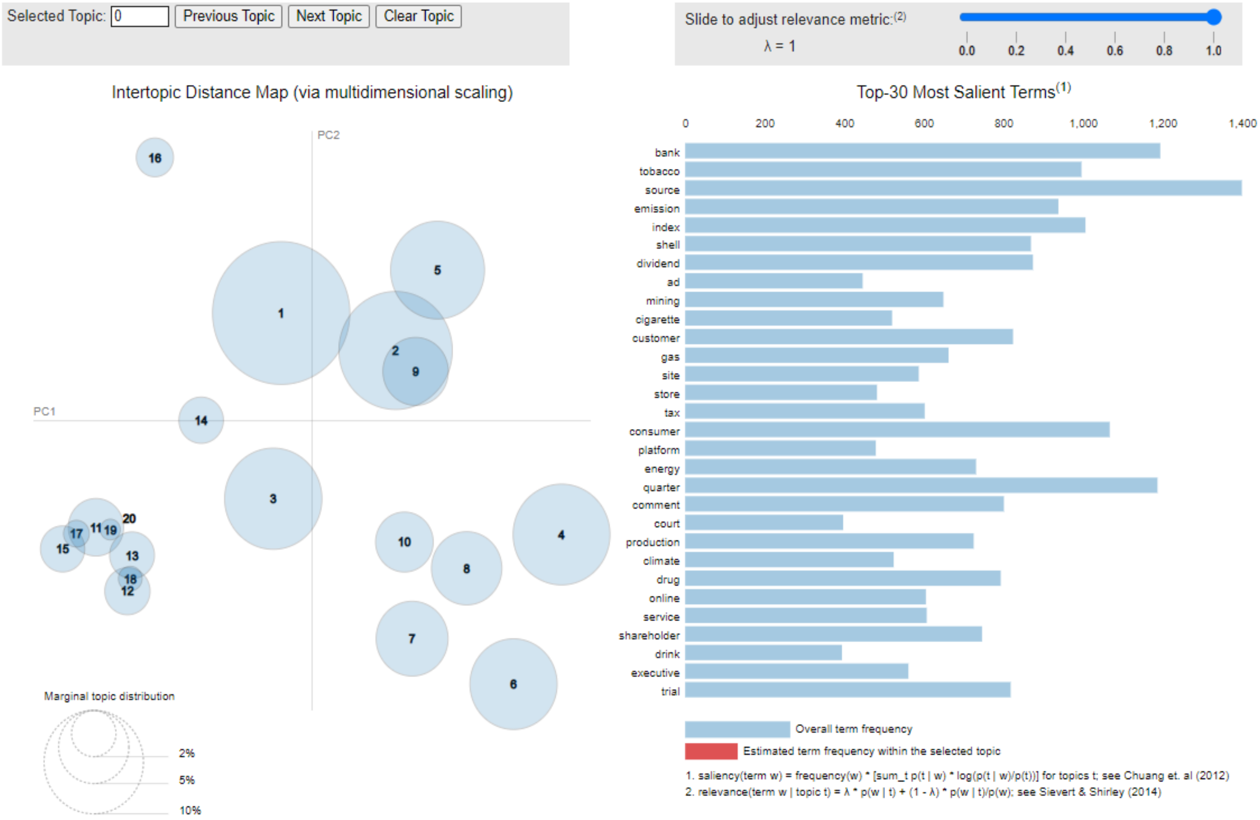
Appendix 3 – Topic Modelling

Commented code is available in notebooks Topic_modelling.ipynb and topic_visualisation FINAL.ipynb.

3.1 K-number of topics sensitivity tests

K	*Coherence score*
2	0.3832
8	0.3819
14	0.4287
20	0.39
26	0.4025
32	0.4174
38	0.3935

3.2 Intertopic Distance Map (left) and Top-30 Most Salient Terms (right)



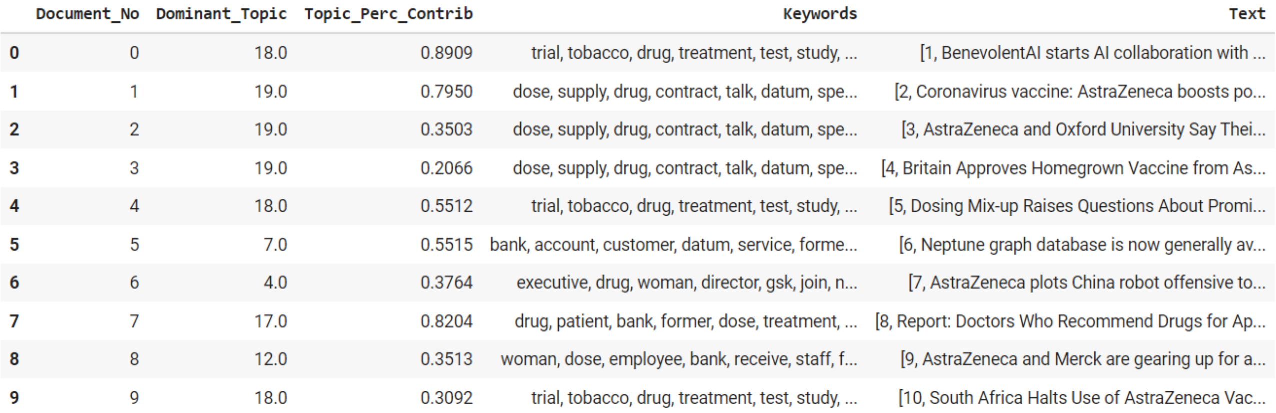
Selected Topic: 0

Slide to adjust relevance metric:[2]

λ = 1

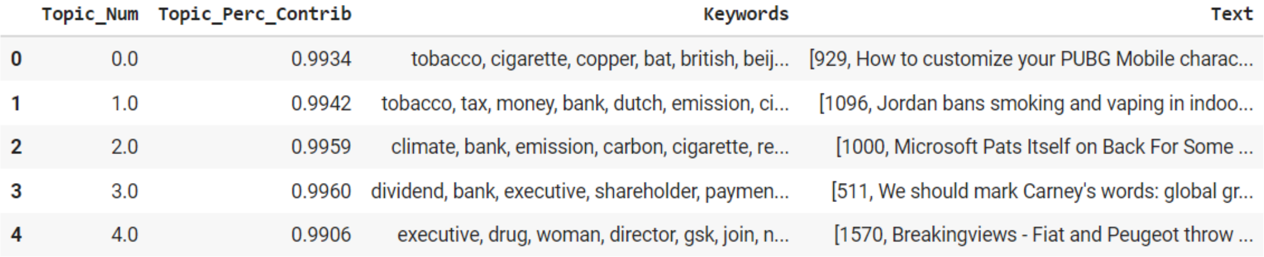
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

3.3 The most dominant topic in sentences. A snippet of the first nine sentences.



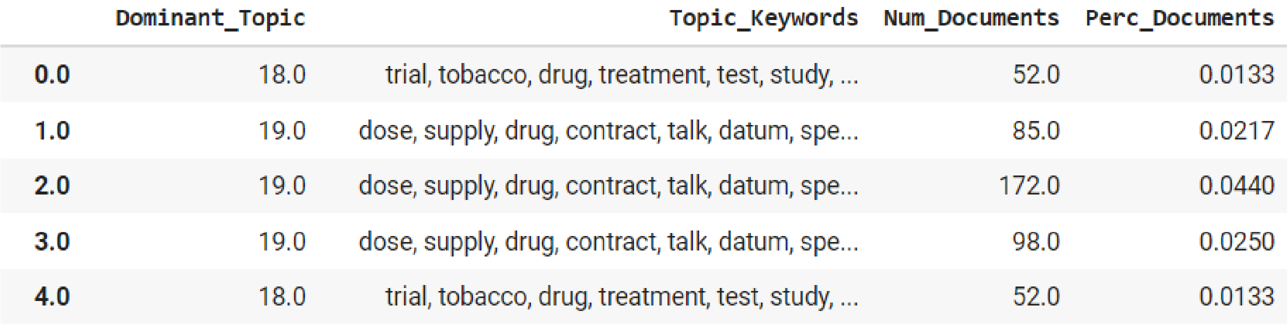
	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	18.0	0.8909	trial, tobacco, drug, treatment, test, study, ...	[1, BenevolentAI starts AI collaboration with ...
1	1	19.0	0.7950	dose, supply, drug, contract, talk, datum, spe...	[2, Coronavirus vaccine: AstraZeneca boosts po...
2	2	19.0	0.3503	dose, supply, drug, contract, talk, datum, spe...	[3, AstraZeneca and Oxford University Say Thei...
3	3	19.0	0.2066	dose, supply, drug, contract, talk, datum, spe...	[4, Britain Approves Homegrown Vaccine from As...
4	4	18.0	0.5512	trial, tobacco, drug, treatment, test, study, ...	[5, Dosing Mix-up Raises Questions About Promi...
5	5	7.0	0.5515	bank, account, customer, datum, service, forme...	[6, Neptune graph database is now generally av...
6	6	4.0	0.3764	executive, drug, woman, director, gsk, join, n...	[7, AstraZeneca plots China robot offensive to...
7	7	17.0	0.8204	drug, patient, bank, former, dose, treatment, ...	[8, Report: Doctors Who Recommend Drugs for Ap...
8	8	12.0	0.3513	woman, dose, employee, bank, receive, staff, f...	[9, AstraZeneca and Merck are gearing up for a...
9	9	18.0	0.3092	trial, tobacco, drug, treatment, test, study, ...	[10, South Africa Halts Use of AstraZeneca Vac...

3.4 The most representative news article for each topic. A snippet of the first four topics.



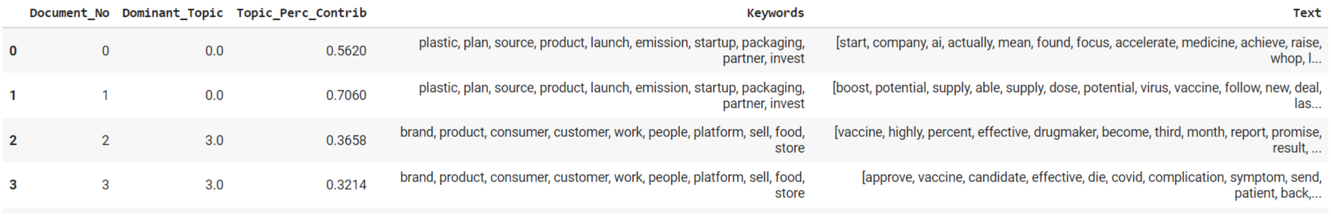
	Topic_Num	Topic_Perc_Contrib	Keywords	Text
0	0.0	0.9934	tobacco, cigarette, copper, bat, british, beij...	[929, How to customize your PUBG Mobile charac...
1	1.0	0.9942	tobacco, tax, money, bank, dutch, emission, ci...	[1096, Jordan bans smoking and vaping in indoo...
2	2.0	0.9959	climate, bank, emission, carbon, cigarette, re...	[1000, Microsoft Pats Itself on Back For Some ...
3	3.0	0.9960	dividend, bank, executive, shareholder, paymen...	[511, We should mark Carney's words: global gr...
4	4.0	0.9906	executive, drug, woman, director, gsk, join, n...	[1570, Breakingviews - Fiat and Peugeot throw ...

3.5 Topic distribution across documents. A snippet of the first four documents.



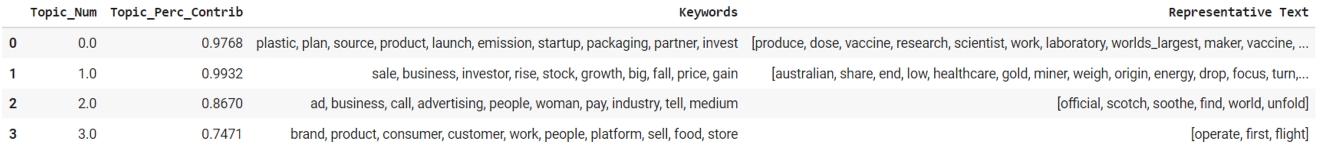
	Dominant_Topic	Topic_Keywords	Num_Documents	Perc_Documents
0.0	18.0	trial, tobacco, drug, treatment, test, study, ...	52.0	0.0133
1.0	19.0	dose, supply, drug, contract, talk, datum, spe...	85.0	0.0217
2.0	19.0	dose, supply, drug, contract, talk, datum, spe...	172.0	0.0440
3.0	19.0	dose, supply, drug, contract, talk, datum, spe...	98.0	0.0250
4.0	18.0	trial, tobacco, drug, treatment, test, study, ...	52.0	0.0133

3.6 Top Four Topics: Dominant topic and its percentage contribution in each document. A snippet of the first three documents.



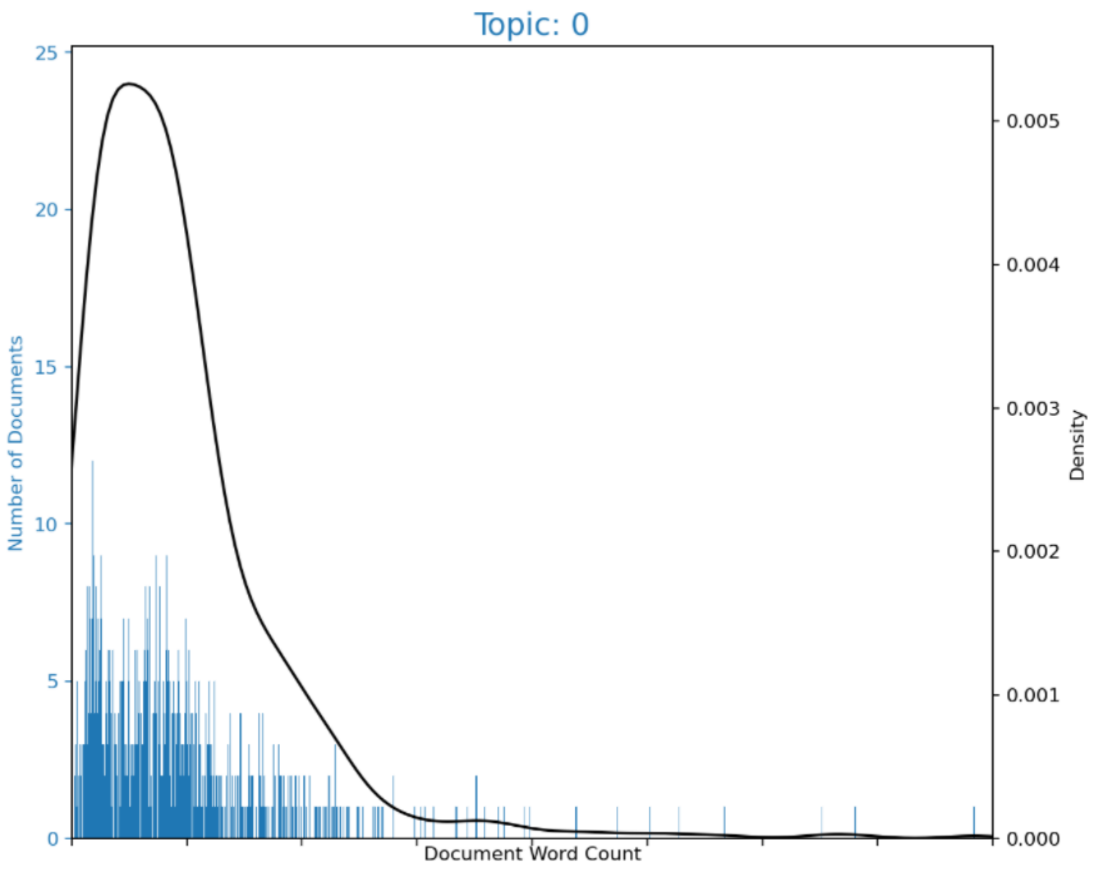
	Document_No	Dominant_Topic	Topic_Perc_Contrib	Keywords	Text
0	0	0.0	0.5620	plastic, plan, source, product, launch, emission, startup, packaging, partner, invest	[start, company, ai, actually, mean, found, focus, accelerate, medicine, achieve, raise, whop, l...
1	1	0.0	0.7060	plastic, plan, source, product, launch, emission, startup, packaging, partner, invest	[boost, potential, supply, able, supply, dose, potential, virus, vaccine, follow, new, deal, las...
2	2	3.0	0.3658	brand, product, consumer, customer, work, people, platform, sell, food, store	[vaccine, highly, percent, effective, drugmaker, become, third, month, report, promise, result, ...
3	3	3.0	0.3214	brand, product, consumer, customer, work, people, platform, sell, food, store	[approve, vaccine, candidate, effective, die, covid, complication, symptom, send, patient, back,...

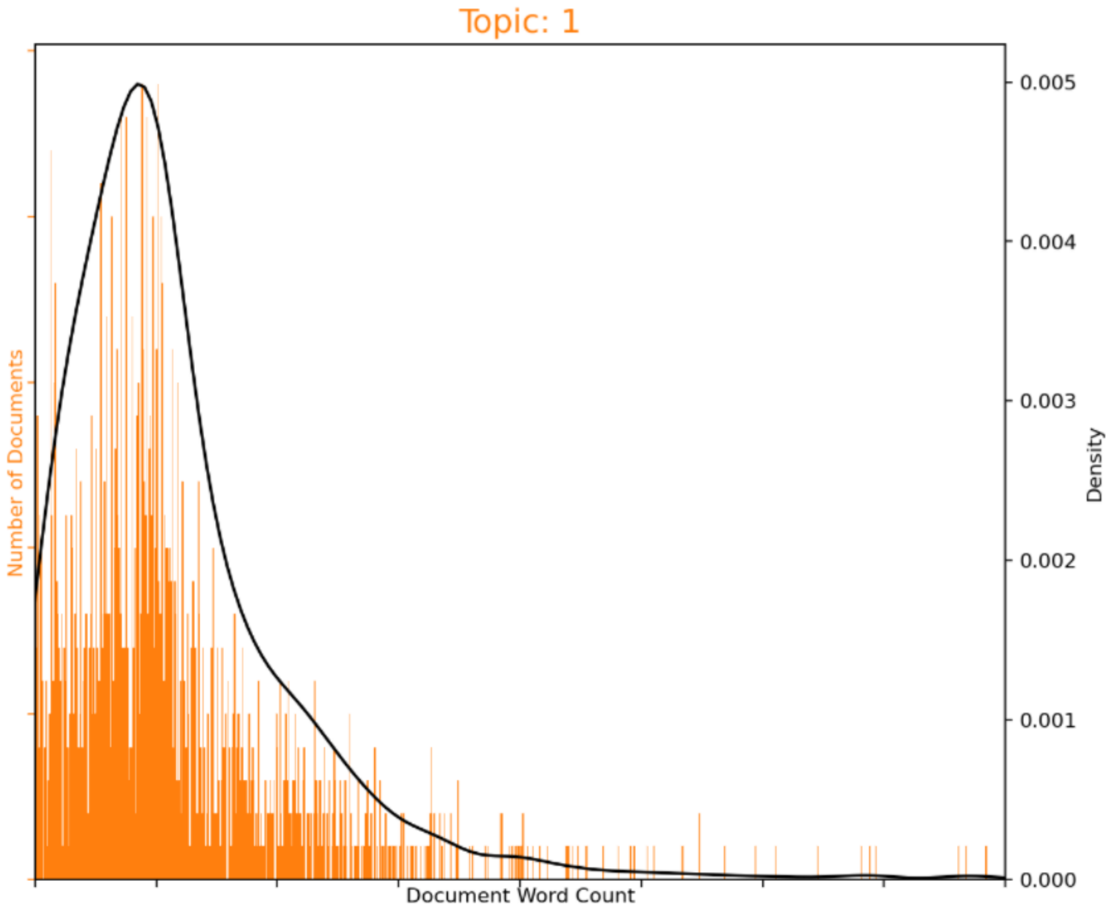
3.7 Top Four Topics: The most representative sentence for each topic

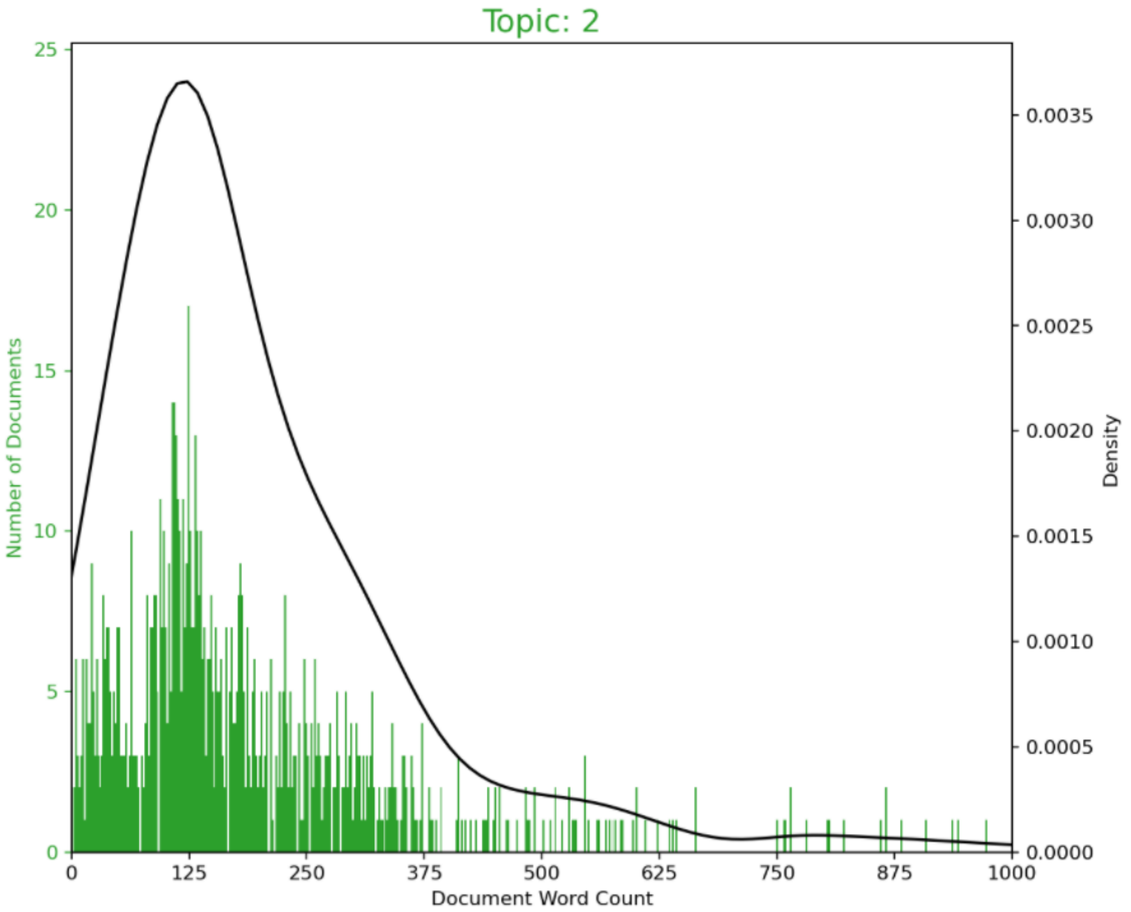


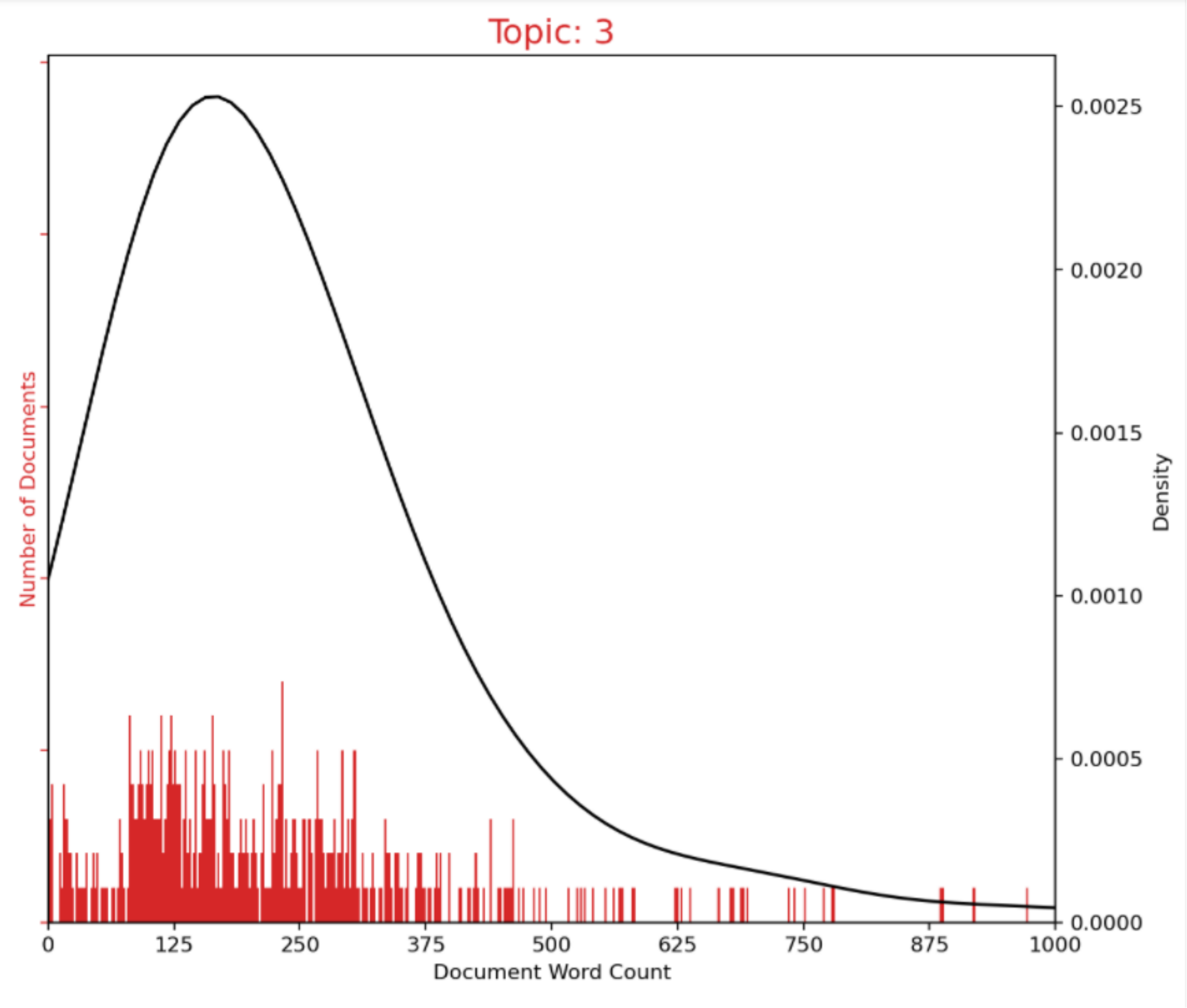
	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0.0	0.9768	plastic, plan, source, product, launch, emission, startup, packaging, partner, invest	[produce, dose, vaccine, research, scientist, work, laboratory, worlds_largest, maker, vaccine, ...
1	1.0	0.9932	sale, business, investor, rise, stock, growth, big, fall, price, gain	[australian, share, end, low, healthcare, gold, miner, weigh, origin, energy, drop, focus, turn,...
2	2.0	0.8670	ad, business, call, advertising, people, woman, pay, industry, tell, medium	[official, scotch, soothe, find, world, unfold]
3	3.0	0.7471	brand, product, consumer, customer, work, people, platform, sell, food, store	[operate, first, flight]

3.8 Top Four Topics: Frequency distribution of word counts in documents

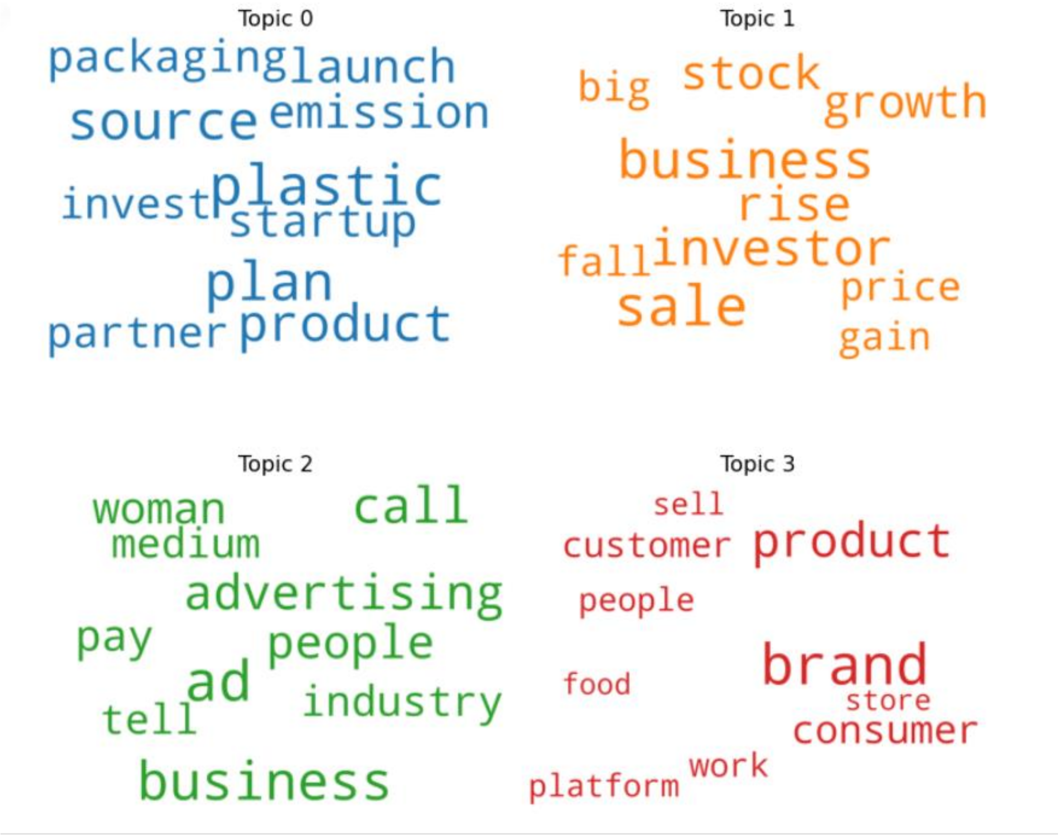



Topic: 1
Number of Documents
Document Word Count
Density
0.005
0.004
0.003
0.002
0.001
0.000

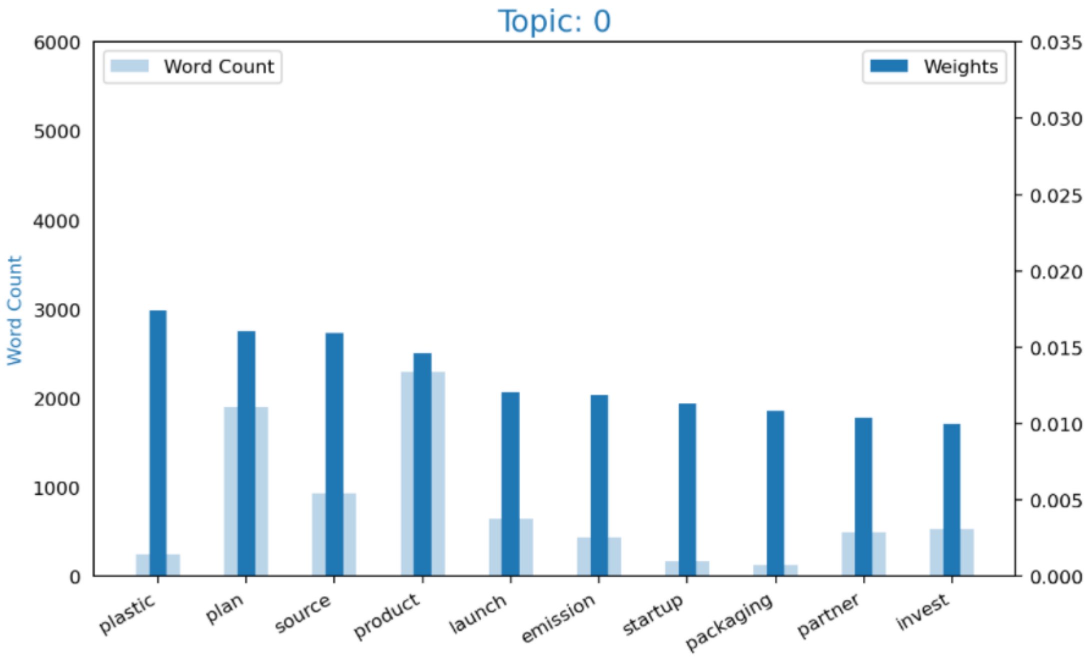

Topic: 2
Number of Documents
25
20
15
10
5
0
0
125
250
375
500
625
750
875
1000
Document Word Count
Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000

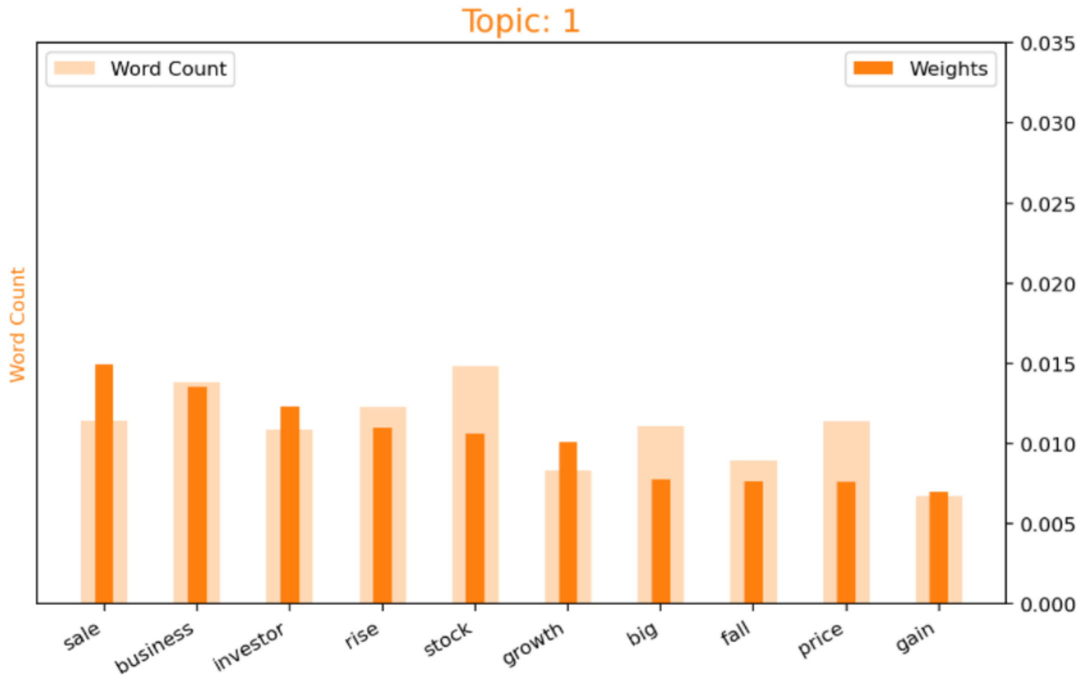

Topic: 3
Number of Documents
Density
Document Word Count
0
125
250
375
500
625
750
875
1000
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025

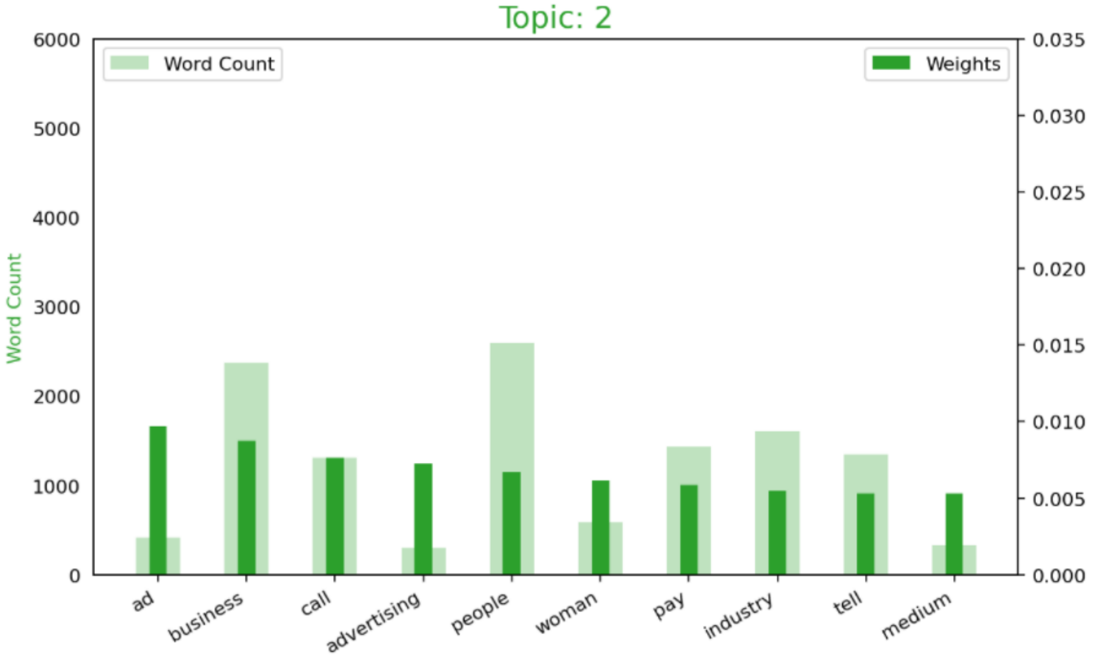
3.9 Top Four Topics: Word clouds of top n keywords in each topic

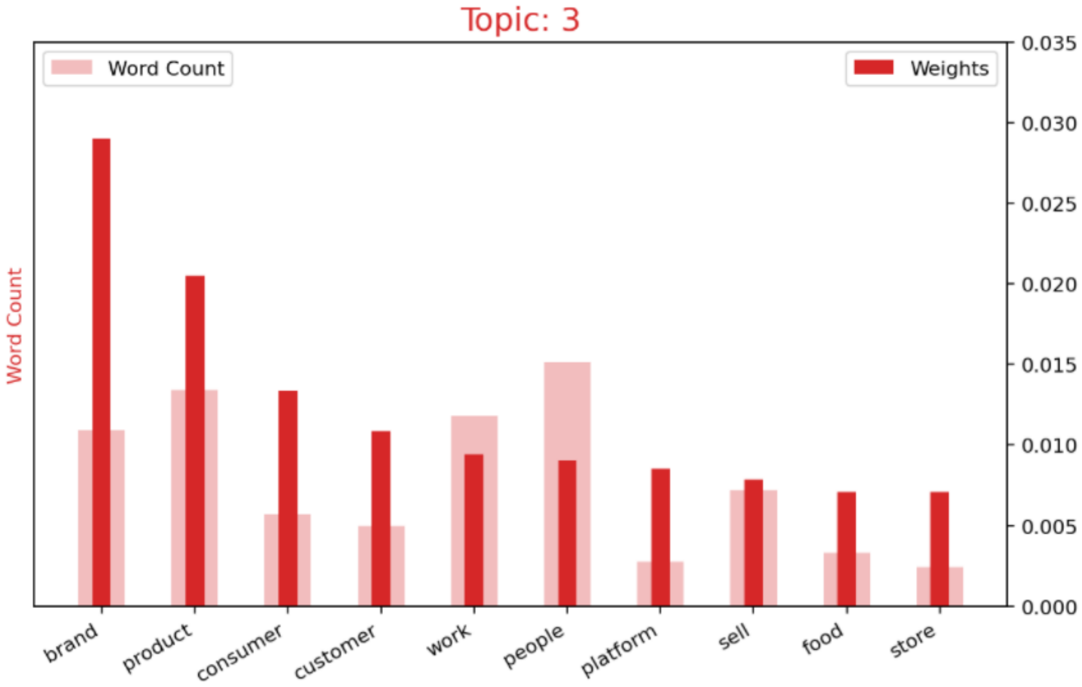


3.10. Top Four Topics: Word count and importance of topic keywords

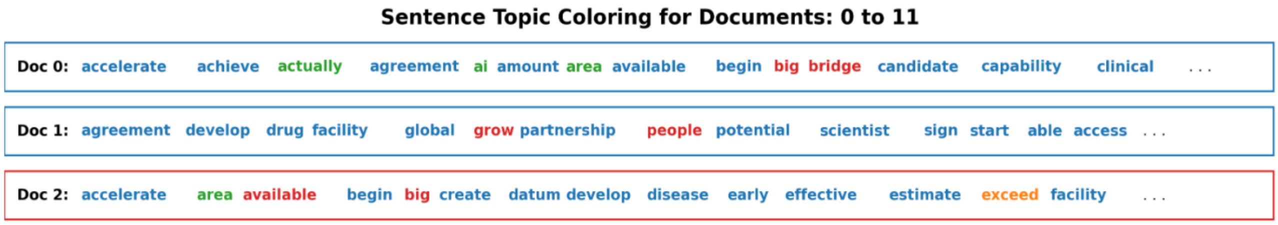



Topic: 1
Word Count
Weights
Word Count
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
sale
business
investor
rise
stock
growth
big
fall
price
gain


Topic: 2
Word Count
Weights
Word Count
6000
5000
4000
3000
2000
1000
0
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
ad
business
call
advertising
people
woman
pay
industry
tell
medium



3.11 Top Four Topics: Sentence chart coloured by topic. A snippet of the first three documents. Each colour corresponds to one of the four topics shown above.



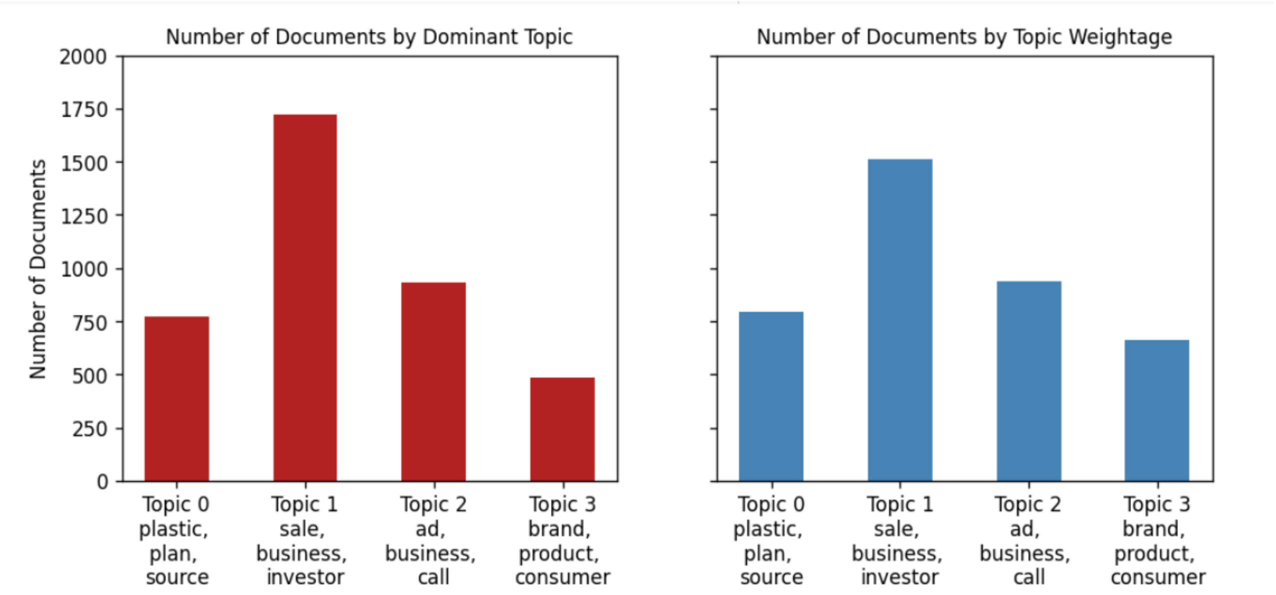
Sentence Topic Coloring for Documents: 0 to 11

Doc 0: accelerate achieve actually agreement ai amount area available begin big bridge candidate capability clinical ...

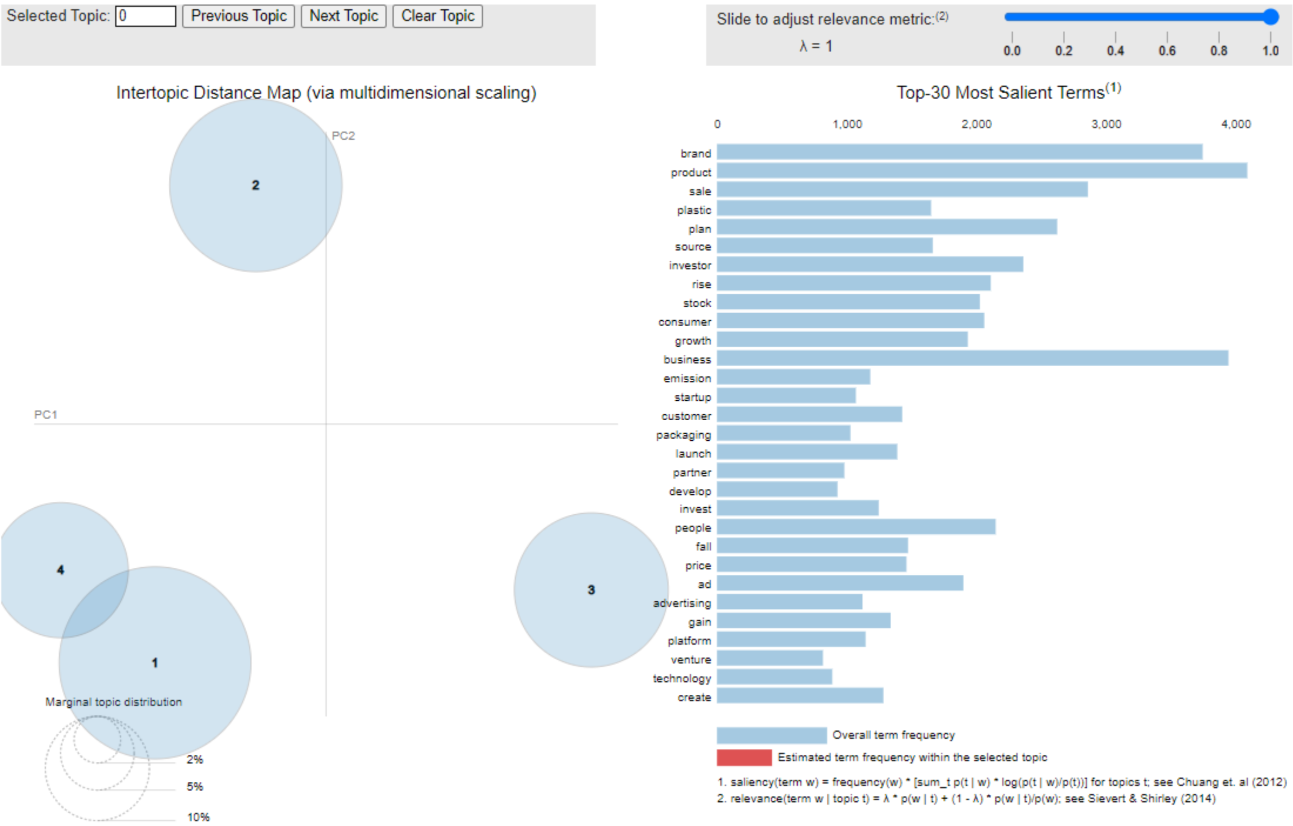
Doc 1: agreement develop drug facility global grow partnership people potential scientist sign start able access ...

Doc 2: accelerate area available begin big create datum develop disease early effective estimate exceed facility ...

3.12 Top Four Topics: The most discussed topics in documents



3.13 Top Four Topics: pyLDAVis interactive chart.



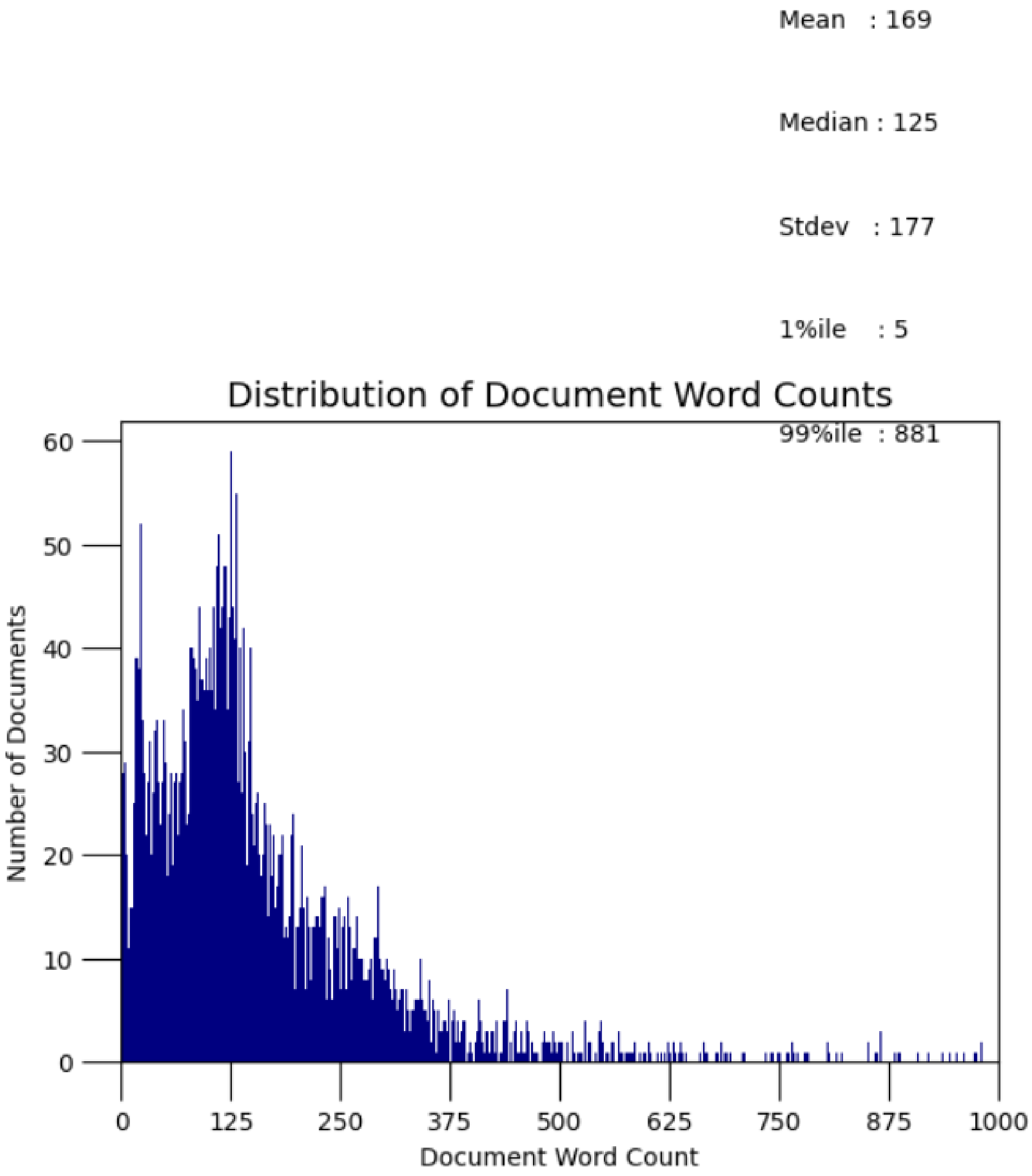
Selected Topic: 0

Slide to adjust relevance metric:(2)
λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

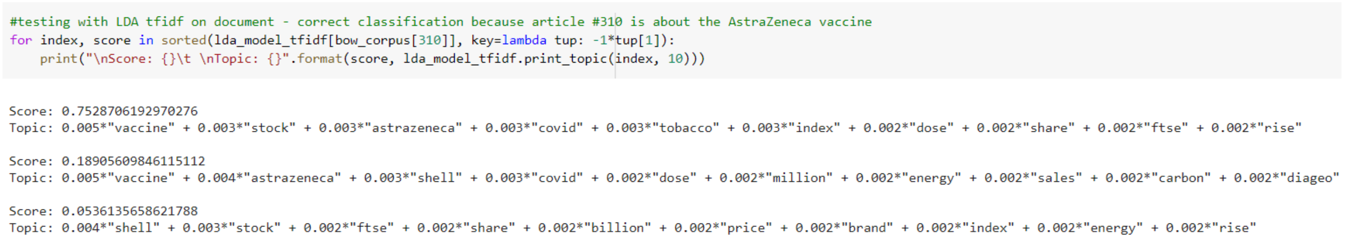
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

3.14 Top Four Topics: Distribution of document word counts



3.15 Topic Modelling for classification: Experiment outputs

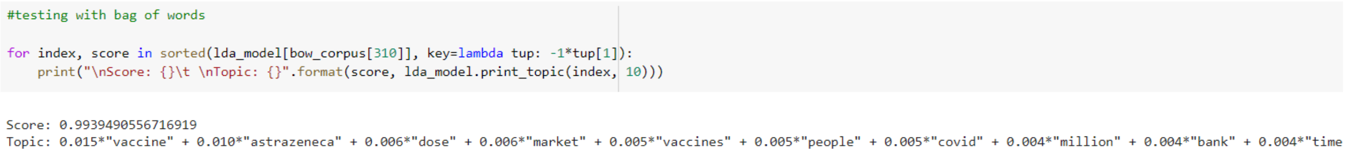
Testing with TF-IDF proved successful.



```
#testing with LDA tfidf on document - correct classification because article #310 is about the AstraZeneca vaccine
for index, score in sorted(lda_model_tfidf[bow_corpus[310]], key=lambda tup: -1*tup[1]):
    print("\nScore: {}\t \nTopic: {}".format(score, lda_model_tfidf.print_topic(index, 10)))
```

```
Score: 0.7528706192970276
Topic: 0.005*"vaccine" + 0.003*"stock" + 0.003*"astrazeneca" + 0.003*"covid" + 0.003*"tobacco" + 0.003*"index" + 0.002*"dose" + 0.002*"share" + 0.002*"ftse" + 0.002*"rise"

Score: 0.18905609846115112
Topic: 0.005*"vaccine" + 0.004*"astrazeneca" + 0.003*"shell" + 0.003*"covid" + 0.002*"dose" + 0.002*"million" + 0.002*"energy" + 0.002*"sales" + 0.002*"carbon" + 0.002*"diageo"

Score: 0.0536135658621788
Topic: 0.004*"shell" + 0.003*"stock" + 0.002*"ftse" + 0.002*"share" + 0.002*"billion" + 0.002*"price" + 0.002*"brand" + 0.002*"index" + 0.002*"energy" + 0.002*"rise"
```

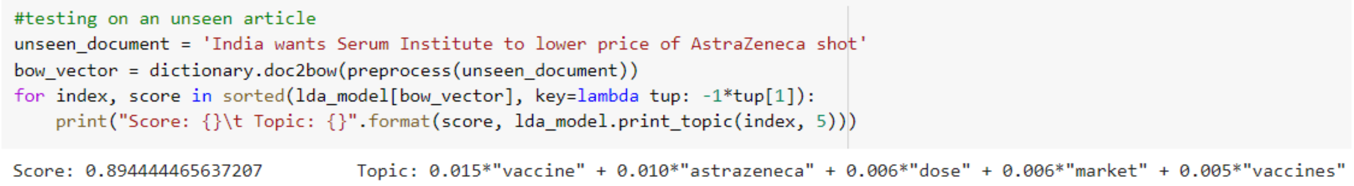
Testing with bag-of-words proved successful.



```
#testing with bag of words

for index, score in sorted(lda_model[bow_corpus[310]], key=lambda tup: -1*tup[1]):
    print("\nScore: {}\t \nTopic: {}".format(score, lda_model.print_topic(index, 10)))
```

```
Score: 0.9939490556716919
Topic: 0.015*"vaccine" + 0.010*"astrazeneca" + 0.006*"dose" + 0.006*"market" + 0.005*"vaccines" + 0.005*"people" + 0.005*"covid" + 0.004*"million" + 0.004*"bank" + 0.004*"time
```

Testing on an unseen article was accurate.



```
#testing on an unseen article
unseen_document = 'India wants Serum Institute to lower price of AstraZeneca shot'
bow_vector = dictionary.doc2bow(preprocess(unseen_document))
for index, score in sorted(lda_model[bow_vector], key=lambda tup: -1*tup[1]):
    print("Score: {}\t Topic: {}".format(score, lda_model.print_topic(index, 5)))
```

```
Score: 0.894444465637207        Topic: 0.015*"vaccine" + 0.010*"astrazeneca" + 0.006*"dose" + 0.006*"market" + 0.005*"vaccines"
```

3.16 Python packages used in topic modelling

Main packages	Numpy, pandas, pprint, nltk, re, sklearn

To enable genism and pyldavis installation	mUsingColab = False
	
	if imUsingColab:
	    !pip install gensim
	    !pip install pyLDAvis
	    !pip install vega
	    !pip install altair

LDA model	Genism, from genism.corpora corpora, from genism.models CoherenceModel, mallet , LdaModel
Evaluation	gensim.models import CoherenceModel, accuracy_score
Pre-processing	from genism.utils simple_preprocess, nltk stopwords, from gensim.parsing.preprocessing STOPWORDS,
Lemmatization	Spacy

Enable logging for gensim	import logging
	logging.basicConfig(format='%(asctime)s : %(levelname)s : %(message)s', level=logging.ERROR)

To remove warnings where possible	Warnings
Train/test	train_test_split
Visualisation	Pyldavis, matplotlib