ESG Classification in Social Media

Yiqing Zhao

MSc Computational Finance

Department of Computer Science
University College London

2022

Supervisor: Caccioli Fabio

Disclaimer

This dissertation is submitted as part requirement for the MSc Computational Finance degree at UCL. It is substantially the result of my work except where explicitly indicated in the text."

The dissertation may be freely copied and distributed, provided the source is explicitly acknowledged.

Or if your project includes information that prevents it from being more widely circulated:

The dissertation will be distributed to the internal and external examiners but thereafter may not be copied or distributed except with permission from the author.

Abstract

The Environmental, Social and Governance (ESG) considerations have gained much attention in investment strategies and business regulatory and marketing decisions in recent years. As a result, more and more corporations spend time on ESG analysis to control the risk. Many financial and non-financial companies are working on ESG rating, which reflects the companies' E-S-G performance. We analysed ESG risk from Tweets data since Tweets provided fast and massive information about a particular company. This dissertation mainly focuses on classifying Twitter data based on ESG class. We use the K-mean clustering algorithm to label the data and a machine learning model (logistic regression, Naïve Bayes and Support vector machines) to predict the classification.

Acknowledgements

Thanks to my team members, thanks to Harvey, Fabio and Joe.

And huge thanks to myself.

List of Figures

List of tables

Chapter 1 Introduction

1.1 Introduction and Rationale

Nowadays, reputational risk analysis has become essential to the risk analysis area. Corporations can control risk from risk analysis since risk analysis identifies and evaluates factors that may negatively impact a business or project's success. As a result, companies can examine the risk of decisions to prepare for adverse events and minimise their effects. As a result of data science and increased amount of data available for analysis, reputation risk is one of the emerging risks that can be analysed and predicted to support investment decisions.

Sentiment analysis uses Artificial Intelligence of natural language processing to study affective states and personal information by systematically identifying, extracting, quantifying, and analysing the data. For example, sentiment analysis could quantify the reputational risk of the company's public comments. Environmental, Social and Governance (ESG) risks is an essential domain for sentiment analysis applied. One of the areas in sentiment analysis is the Environmental, Social and Governance (ESG) risks. America's Business Roundtable CEO released a Statement on the Purpose of a Corporation (SPC) on August 19, 2019. The SPC stated that companies should not only consider the benefit of shareholders but also need to send values to customers, invest in their employees, consider the environment and handle the relationship of the suppliers and the communities they operate. This new statement marked a new modern standard for companies' responsibility that corporations should pay more attention to Environmental, Social and Governance considerations. Moreover, by the end of 2021, ESG information was used in over 120 trillion dollars in signed agreements for investment decisions [1]. This shows that the market considers ESG information when they process risk control.

The ESG risk could cause reputation risk as ESG represents the company's performance in three layers. First, ESG rating has become a vital component of capital reallocation, so these products' market demand and supply are only expected to grow [2]. To quantify this risk, we need a large data basis to do the sentiment analysis, which is essential for gathering data. The introduction of social media

presents a massive database; compared to the old time, the new technology developed a more accessible, cheaper and faster platform for people to share the news. The social media platform can provide the latest information about companies.

Moreover, with the advent of social media, reputation events are impacting stock prices twice as much [3]. This indicated that social media data could be reliable for reputation risk analysis. Given the close relation between ESG risk and reputation risk, we can use machine learning and Natural Language Processing methods to analyse the reputation risk in E-S-G three aspects. However, before detecting the ESG scores, we need to classify the large dataset we gathered from social media into three Environmental, Social and Governance classes.

1.2 Research Objectives, Questions and Techniques

My work focuses on the social dimension of ESG. The social risk is related to employee wellbeing, health and safety, labour laws and human rights. In statistics, classification is the problem of determining which category an observation belongs to. I will use in particular data collected from Twitter, and the goal is to distinguish between tweets that discuss a social problem vs. those that do not.

Moreover, the dissertation is structured as follows: In chapter 1 and 2 I will discuss how sentient analysis can play a role in risk control and why ESG is a vital risk for financial decision making. Furthermore, I will discuss the concept of classification and how this can be applied to risk analysis. In chapter 3 I explained all the technology that related to my research. Chapter 4 and 5 gives the result of the project, with explanation of the model performance and determination of the social related event time.

Our data are related to some social events, for instance, the Ryanair racism incident, during a Ryanair flight FR9015 in 2018, an old black woman seated next to a white man was racially abused.[37] We classify our data that make it as two clusters, the first is an event-related cluster, and the other is a non-related cluster. Moreover, we

labelled the data as social-related (for event-related clusters) and non-social-related data. Furthermore, use these data to train a supervised machine learning model, i.e., Logistic regression, Naïve Bayes and SVM. See how the model performed.

1.3 Initial Hypotheses

We are using the data from Twitter to build a social-related classification model. We are a reputation advisor that observes reputation changes. The classification is used to see how the companies handled ESG issues, and it can break down the reputation of ESG and further details. And the model can be used to classify other tweets. The task is to evaluate the performance of the model.

1.4 Research Methodology and Paper Structure

Kennedys have processed the data I received. The data is related to 7 companies' news that affects the companies in E-S-G three layers. For example, in 2018, a teenager died after eating a sandwich containing sesame, which caused an extreme allergic reaction. [36] There were no legal obligations to specify the label of Pret's product in this presence. This causes some furore on social media and may bring social-related reputation loss to Pret. Kennedys calculated the sentiment score of each tweet. The time range of each company's data is different since the event time is different. So that our analysis is mainly focused on the event's consequences, the details of all data will be introduced in Chapter 3.

The way I am going to classify the tweet with also be introduced in Chapter 3. First, we need to process the text before processing the classification algorithm. Next, we should eliminate most noise, e.g., remove the emoji, the URL link, and the symbols. Then we remove frequently used words and common suffices. After all the processing, we left with the essential part, and we labelled the data with TF-IDF clustering so that we could train a supervised machine learning model. Furthermore finally, test your model to see how the machine learning is going.

Overall we have the result in chapter 4, which mainly indicates the performance of our model and I showed that the K-mean cluster can distinguish between tweets related to the event and not related to the event. This can be used to create a labelled dataset, which I used to trained a classifier that can distinguish between two classes. And then, we have the conclusion in chapter 5 with an indication of further works. Because we only have limited time and data, chapter 5 shows what we should do if we have more time and data.

Chapter 2 Literature review

In this section, I presented the related work on the ESG rating market, sentiment analysis and different use of Twitter data. The ESG rating market is currently in its nascent state, with inaccurate regulation and unclear standards of rating measures. [6] Each rating company has their rating system and standard. Over the past decade, ESG ratings have gained increasing importance to investors, who use them to improve a company's societal impact. [9-13] The rapidly increasing interest in ESG rating creates massive demand for ESG data. [6,7] There were already more than 600 ESG ratings and rankings globally in 2018; with the exploding increase of the market size, regulations are necessary for the market. The barriers in the ESG rating market are the quality, consistency, and reliability of ESG data, which are the components investors consider while choosing ESG rating companies. [2,15]

However, there is research in 2020 finding that compared to the correlation between investment decision and credit rating (0.99), the correlation of ESG rating is much lower (0.38). [2,4,14] The gap is probably caused by three factors. The first is that ESG performance is less likely to impact bonds or shares. Second, companies may be confused by mixed signals from different ESG rating agencies. Finally, financial institutions struggle to reflect ESG profiles in their disclosures and pricing properly. [2,6]

The ESG rating market used to face a lack of direct regulation. [8] Nevertheless, with international and national announcements, the regulation of the ESG rating market will begin in late 2022. According to HMT's 'Greening Finance: A Roadmap to Sustainable Investing, published in October 2021, the UK government considered digitising ESG data, potentially including a centralised register, as being considered by the government and regulations.

The Financial Conduct Authority (FCA) released its ESG strategy in October 2021; the report addressed two critical issues in the ESG rating market, transparency and trust. [2] For clarity, there are several aspects to consider: checking the data source

and data quality, restricting the use of unreliable data sources, and revealing the methodology in detail, including the use of estimations and algorithms, governance on changes made to the method, clear definition of the rating objective, and interaction with companies that will be rated. Through the delivery of ESG-labelled securities, products, and services, trust is demonstrated in the effective integration of ESG into financial market decisions. [2,15,16] As mentioned in the last paragraph, this is in line with the third factor.

From April 2022, the UK government required companies and financial institutions with more than 500 employees and over £500m in annual turnover to disclose climate-related financial data on a mandatory basis by the task force on climate-related financial disclosures (TCFD). As a result, the UK will become the first G20 country to enshrine the mandate in law. UK officials are expected to adopt the newly published ISSB standard. [2] By 2023, the UK government will regulate other UK-authorised asset managers, life insurers, and pension providers regulated by the FCA. The TCFD disclosures will also be applied to all other occupational pension schemes in 2024-2025.[5]

For sentiment analysis, the idea is to use sentiment score as a reputation index that enables the companies to respond to feedback quickly and develop powerful market strategies. In recent years, there have been many ways to measure reputation since the increased use of internet data. [17,18] The problems of measuring reputation loss are, first, what precisely makes the ‘reputation’ stands for? Second, how to quantify it? In [19], Deloitte has some approaches to these problems. Some companies determined the reputation loss by the survey. [20] Mitic Peter gives some arguments about the problems. See [21,22] Mitic gives specific definitions for sentiment and reputation, and [23] gives a complete discussion about how a score is derived. First, we have to ensure that the score is in the range [-1,1]. Although the range is arbitrary, it seems intuitive to cast the score as positive and negative. Mitic’s definition of the reputation of organisation G at time t is:

$$R_G(t) = \frac{\sum_{i=1}^{n} w_i s_{i;H}}{\sum_{i=1}^{n} w_i} \quad (1)$$

Here w_i is the weight of comment i, which are arbitrary but corresponds with the holder score $s_{i;H}$. The holder score $s_{i;H}$ is denoted by the corresponding score with H representing the holder (an agent who makes comment at time t), i denoting same element as in w_i. Here equation (1) illustrates the reputation index only at a single time t. More generally, we create a time series $\hat{R}_G$. Which defined in equation (2)

$$\hat{R}_G = \{R_G(t)\}_{t \in T} \quad (2)$$

The processes of reputation index calculated explained in 5 steps:

1. "Comments" are received electronically from opinion holders, H, that convey sentiment regarding a target G (news channels, social media, etc.).
2. Each comment processed with sentiment analysis, resulting in a sentiment "score" nominally in the range [−1, 1].
3. Determine a weight w_i for each comment (e.g., to reflect the opinion holder's influence).
4. Utilizing all sentiment scores received over a given period (such as one day), compose a reputation index applicable at that time. (use equation(1))
5. Form the reputation time series $\hat{R}_G$ by accumulating successive reputation indexes over time (use equation(2))

In Chapter 3, I will introduce TF-IDF which is similar to the w_i it is a way to calculate the weight of some words. I will use TF-IDF to do the data labelling. In [17], Mitic shows that companies' reputation, i.e., sentiment score, has an impact on the company's financial status. So that it is valuable to do the sentiment analysis in order to control the emerging risk.

As Twitter data is seemingly readily accessible, easy to access, and inexpensive (at least on a small scale), it is widely seen as a valuable source for social, Information and Communication Technology (ICT) and other research. Although it is controversial to use Twitter data since there are some potential issues, such as ethical issues and legal constraints on the collection and use of Twitter data, [28] Twitter data still play essential roles in the research area. In [24], Tweeter data has been used to measure customer satisfaction and predict customer churn. Moreover, Tweets have also been used in Disaster Management [29-32], disaster identification [33], waste minimisation [34] and modelling stock volume [35]. The primary use of tweeter data is that prediction (mentioned in this paragraph) and sentiment analysis (mentioned in the next paragraph).

For more microblogging brand impact analysis, a comparison of Twitter and other social networks see [26], indicating that Twitter has the potential to become an vital application in the economic and financial area. In [25], Their approach involves automatically classifying Twitter messages based on their sentiment. Since Twitter is already a popular microblogging service where users can post status messages (also known as "tweets"), their study classified tweets into 177 negative tweets and 182 positive tweets with strip emoticons out of the training data and used Naive Bayes, Support Vector Machines, and maximum entropy as machine classifiers. Approximately 80% to 83% of the samples were accurate.

In later 2010, [27] tried a more extensive data set with 30,000 tweets as a corpus, they train a sentiment classifier using the collected corpus. They used N-grams as features in their Naive Bayes multinomial classifier and achieved their best results with the NB classifier. So that in this report, I'm going to use Logistic regression, NB and SVM as the supervised machine learning model for classification.

Chapter 3 Methodology

Kennedys Law collects the data we receive, and it is processed initially. So we have got seven companies' tweet data; they have the following columns:

- _id
  - Unique identifier from our DB
- Type
  - Type of data, in this case it will only be tweet
- tweet_id
  - Twitter id of tweet
- Date
  - Datestamp of tweet
- user_name
  - Username of the tweeter
- Text
  - Full text of the tweet
- Sentiment
  - Calculated sentiment using the VADER algorithm

The sentiment score included is produced by Kennedys Law. What we are interested in is the 'text' part for classification. I've been told that the data is a time series where the event happened in the exact middle range; I will test this using the change of sentiment score. The process of classification is done in the following way.

(i) Collect related twitter data of each company.

(ii) Pre-processed the collected data for further analysis.

(iii) Clusters detected using K-Mean clustering algorithm.

(iv) Use labelled data to train a supervised machine learning model.

(v) Test the model and see its performance.

For each companies' data, it is related to the specific events shown in Table 1.

Company	Event	reference
Pret A Manger	A teenager died after eating a sandwich containing sesame in 2018, due to an extreme allergic reaction. This is because the label of the product had not been specified in the presence of this, because of no legal obligations.	[36]
Ryanair	During a Ryanair flight FR9015 in 2018, an old black woman seated next to a white man was racially abused.	[37]
Exxon Mobile	As a result of Hurricane Harvey, two refineries owned by ExxonMobil were damaged and hazardous pollutants were released.	[38]
Ted Baker	Ted Baker's chief executive Ray Kelvin stepped down after over 200 employees demanded that "forced hugging" end.	[39]
Rio Tinto	In 2021, a mining giant Rio Tinto destroyed two sacred rock shelters in Western Australia's Pilbara region	[40]
Tesco	Horsemeat is present in 29% of Tesco's beef-feathered burgers.	[41]
Volkswagen	The Volkswagen Group received a Clean Air Act violation notice from the United States Environmental Protection Agency (EPA).	[42]

Table 3. Indication of events of 7 companies' data

3.1 Natural Language Processing

In terms of communication among individuals, language plays a very significant role. It differentiates the human being from other living creatures. Generally, language (both written and verbal) carries a great deal of information. Whenever we speak or write something, it indicates a topic containing words, grammatical rules, tone signals, etc. Every part of a language provides some information. Analysing the data may result in a combination that suggests some actions.

While working with machines, it is hard to let machines learn about the information in the language. All the data (tweets in this project) we have are unstructured data (i.e., the machine cannot fully understand and process). Ela Kumar defined Natural Language Processing in his book as "Natural Language Processing is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. Natural language generation systems convert information from computer databases into readable human language" [43]

In this report, I will use some techniques of NLP and apply the NLTK library. [45] Natural Language Toolkit, or NLTK, is a set of libraries that perform linguistic natural language processing on English written in the Python programming language. The system was developed by Steven Bird and Edward Loper at the University of Pennsylvania's Department of Computer and Information Science.

The processes state in the following subchapter 3.1.1, 3.1.2, 3.1.3, 3.1.4, and 3.1.5.

3.1.1 Clean the data

The original data consists of 7 columns, whereas in this report, we will only need three columns: date, text and sentiment. So, first of all, I cleaned the data removing unrelated columns. Figure 1 shows the example of the Pret data.

	date	text	sentiment
0	2019-03-30 23:04:23	@Pret Duck Hoisin Wrap is £3.50 out £4.50 in?	0.0000
1	2019-03-30 21:48:41	@Pret More vegan baguettes plz😁	0.1440
2	2019-03-30 21:05:14	@JimmyMc_G @Pret @BBCWatchdog would love all t...	0.6369
3	2019-03-30 19:53:26	.\n\nFREE COFFEES in PRET A MANGER\n\n>>...	0.6166
4	2019-03-30 19:48:48	No, the floor of a train isn't a bin... 😡 @sta...	0.0000
5	2019-03-30 19:32:48	FREE Coffees in Pret A Manger —-> #FREE #Pr...	0.6166
6	2019-03-30 19:07:05	@Pret The vegan cookie which needs to make a c...	0.0000
7	2019-03-30 18:45:48	Strange but lovely experience yesterday; happe...	0.8896
8	2019-03-30 18:31:30	@Pret I have chatted to the team many times, s...	-0.5002
9	2019-03-30 18:27:31	@Pret Thanks at Waterloo Stn to staff for the ...	0.7901

Figure 1 Example of Pret data

We can see lots of punctuation, emoji and URL link that we want to remove so that the data can only show the machine the critical information. The emojis may consist with some information (for example it may related to sentiment), for just for simplicity we dropped them in this analysis.

I did the following steps to clean the data:

1. Drop duplicate data.
2. Lowercase all the data (make it easy for machine learning).
3. Remove the emoji, URL and unrelated characters.

3.1.2 Tokenization

The tokenization technique is straightforward and breaks the sentences/texts into pieces called tokens. It is a pretty basic technique in NLP. In the NLTK library, we can use nltk.tokenize().

3.1.3 Remove stopwords

The stopword is a word such as "and", "the", or "to" in English, which are all widespread words that provide little or no value to NLP objectives, so they need to be filtered or excluded.

3.1.4 Stemming

This process removes the affixes (lexical additions to the word's roots). For instance, "helpful" will become "help". There is a problem with affixes because they can create new forms of a word (called inflectional affixes) or even create new words themselves (called derivational affixes). The English prefixes are always derivatives (they create a new word, such as the prefix "eco" in the word "ecosystem"). In contrast, suffixes are either derivational (the suffix creates a new word as in the case of "guitarist" and "is" in the word "guitarist") or inflectional (the suffix creates a new word form as in the case of "faster"). However, stemmers are easy to use and run very fast. We prefer this method for the hardware we have since it will cause fewer problems. We use nltk.stem.porter.PorterStemmer() to do the steaming process.

3.1.5 Remove the background data

Now, after processing, we got a pure corpus with essential information. As the project leader suggested, we could remove the background data (i.e., the non-event-related data) by deleting all the corpus that occurs before the event. The problem is how we determine the exact time for the event. In this project, I test it with the change of sentiment score and assume the event happened at 00:00 that day. For the result please see chapter 4.1, figures 2-5

3.2 Text Embedding

As a type of representation, word embeddings allow words with similar meanings to be represented similarly. We could use text embedding to process the data further. In

natural language processing, the embedded text represents words and documents and is considered one of the critical breakthroughs of deep learning. In [44], they state, "One of the benefits of using dense and low-dimensional vectors is computational: most neural network toolkits do not play well with very high-dimensional, sparse vectors. … The main benefit of dense representations is generalization power. If we believe some features may provide similar clues, it is worthwhile to provide a representation that can capture these similarities."

A word embedding is an algorithm representing words as real-valued vectors within a predefined vector space. The technique is often called deep learning because it maps each word to a vector and learns the vector values like a neural network. A real-valued vector represents each word, often with tens or hundreds of dimensions, as opposed to sparse word representations, such as one-hot encoding, which require thousands or millions of dimensions.

Distributed representations are learned based on word usage. Consequently, words with similar meanings can have similar representations, capturing their meaning naturally. In contrast, a bag of words model has a crisp but fragile representation unless explicitly managed. Words have different representations regardless of their usage.

3.2.1 TF-IDF Vectorizer

For every word, the Term Frequency Inverse Document Frequency (TF-IDF) value increases with every appearance of the word in a document. However, it gradually decreased with every appearance in other documents. TF-IDF can determine which keywords we used for the clusters in the classification.

We could understand the mathematical concept inside tf-idf:

Term frequency (TF) represents by equation (3):

$$tf(w,d) = \log\left(1 + f(w,d)\right) \tag{3}$$

Here N is the number of documents (Tweets) we have, d is the number of given documents from our data, D is the collection of all documents, and w is given the word in a document. Moreover, f(w,d) represents the frequency of word w in document d. Noted that the word w has to be non-stopwords. (i.e., the word that contains information)

Inverse term frequency (IDF) represents by equation (4):

$$idf(w, D) = \log\left(N/f(w, D)\right) \tag{4}$$

TF-IDF represents by equation (5)

$$tfidf(w, D, d) = \mathrm{tf}(\mathrm{w}, \mathrm{d}) * idf(w, D) \tag{5}$$

The consequence of equation (5) is that a high weight for the tf-idf calculation is reached if the document has a high term frequency(tf) and the document in the collection has a standard term frequency. In python, we could use library sklearn.feature_extraction.text import TfidfTransformer and TfidfVectorizer. Then we could use Tf-idf to do the k-mean cluster in order to label the data.

3.2.2 Count vectorizer

In Python, CountVectorizer is one of the great tools provided by Scikit-Learn. Counting the number of times each word appears in the entire text can turn a given text into a vector. It is useful when we have multiple such texts and wish to convert each word into a vector (for further text analysis). It is also a great way to use text classification. But TF-IDF is definite a better idea for vectorising data since td-idf considers not only the times that appears for each corpus but also considers overall documents of the weight of the words.

3.2.3 Principal component analysis (PCA)

PCA is a technique to reduce the dimensionality of the feature space. It creates new features (the principal components) that are a combination of the original features and are ranked in terms of their contribution to the variance of the dataset. Technically it is defined as a set of points in a real dimension space which contains a sequence of p unit vectors, and the ith vector corresponds to the direction of the

best-fitting line to the data, and the the i-th vector is orthogonal to the previous i-1th vector. If we disregard the less important terms, we eliminate the elements we care less about but preserve the main trends with the highest variances (largest information) in PCA.

3.3 K-mean Cluster

From TF-IDF, we can determine the weight of each token in an entire observation. Then we applied the k-mean algorithm to the vectorized data set. The K-means clustering algorithm is a popular unsupervised machine learning algorithm. A typical unsupervised algorithm makes inferences from datasets using only input vectors without referring to labelled outcomes.

First of all, what is clustering? A cluster is a collection of similar patterns in the data that allow the data to be aggregated into groups (also known as clusters). K-means aims to discover underlying patterns by grouping similar data points together. The K-means algorithm seeks out a fixed number (k) of clusters in a dataset. In this report, we need k=2 since we only need clusters related to social or not, so it is a binary classification. The benefit of the k-mean cluster algorithm is that it minimizes the number of centroids by identifying k-centroids and allocating every data point to the nearest cluster. There are some properties of clusters. First, the data in the same cluster should be similar. Secondly, the data points in the different clusters should be as different as possible.

In order to reach the first property, we need to introduce the term inertia, which is a definition in physics; it means the property of matters that remain in uniform motion in the same straight line unless there is an external force. In this algorithm, inertia is calculated as the sum of the distances between all the data points in the cluster and the centroid of this cluster. And the inertia should stay as low as possible so that there are strong connections between all the data points in this cluster. The inertia decreases as the k number increase.

Data labelling is a technique used in machine learning to identify raw data (images, text files, videos, etc.) and add one or more meaningful tags of information to provide context so that machine learning models can learn from them. For example, a label may indicate a photo containing a bird or a car, what words are pronounced in a recording, or whether an X-ray contains a tumour. Data labelling is required for various use cases, including computer vision, natural language processing, and speech recognition. Currently, most machine learning models rely on supervised learning, which maps inputs to outputs using algorithms. Supervised learning requires labels for the model to learn from to make the right decision. Humans are usually asked to make judgments about unlabelled data in data labelling. For example, a tagger might add tags to all the images in the dataset that have the answer "Yes" to the question "Does the photo contain a bird?". A label can be as coarse as a simple yes/no, or as okay as identifying the pixels associated with a bird. "Model training" uses human-supplied labels to learn patterns behind machine learning models. As a result, new data can be predicted using the models thus trained.

We use sklearn library, sklearn.cluster.KMeans, fit with vectorized data, with k set to 2. After the k-mean clustering, the data will be labelled as 0 and 1, with 0 representing social-related data, and 1 being non-social-related data. I list the top 10 elements in each cluster in Chapter 4.

3.4 Supervised Machine Learning

After we labelled all the data, we got into the following steps: use the labelled data trained in supervised machine learning so that we can use the model to do the classification when we get new data. Supervised Learning is training an algorithm with labelled data and finding the process that best describes the input data in the end (i.e., for a given X makes the best estimation of y (X -> y)). Then, in order to forecast the output values for new data based on the relationships that the supervised learning algorithms have learnt from the initial data sets, they attempt to

model the relationships and dependencies between the target prediction output and the input features.

We are using three typical supervised learning models as the classification prediction model: logistic regression, Naïve Bayes and Support Vector Machines.

3.4.1 Logistic regression

Logistic regression is a supervised machine learning algorithm that can be used for classification, and it is used to predict the probability of categorical dependent variables. This indicates that logistic regression is a suitable model for our classification project. We are doing a binary problem with a binary dependent variable (two clusters of variables). This fits binary logistic regression. However, logistic regression requires a large sample size, and the data size we have might not look good enough.

A logistic regression model can be compared to a linear regression model. However, a logistic regression utilizes a cost function that is more sophisticated and can be described as either the sigmoid function or the "logistic function" rather than a linear function. The cost function is usually restricted to the range of 0 and 1 according to the logistic regression hypothesis. Because it can have a value larger than one or less than 0, which is not feasible according to the logistic regression hypothesis, linear functions cannot accurately represent it.

Sigmoid function (function (6)) converts expected values to probabilities. The function converts any real value between 0 and 1 into another value. We apply the sigmoid function in machine learning to convert predictions to probabilities.

$$f(v) = \frac{1}{1+e^{-v}} \tag{6}$$

Here *v* normally represented a linear function of variables (i.e.; linear regression formula of the hypothesis function (7))

$$\beta_0 + \beta_1 \vec{X} \tag{7}$$

Where X is a vector. And also, we have a combined expression of functions 6 and 7, which is the representation of the logistic function (function (8))

$$P\left(y^{(i)} = 1 | x^{(i)}; \beta\right) = \frac{1}{1+e^{-\beta_0 - \Sigma_{j=1}^{p}\left(\beta_j x_j{}^{(i)}\right)}} \tag{8}$$

We use this function to predict the probability and then create the regression.

3.4.2 Naïve Bayes

The Bayes Theorem shown in function (9) is the foundation of the probabilistic machine learning method known as Naive Bayes, which is utilized for various classification problems.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{9}$$

P(A|B) means the probability of A occurring given that B has already occurred. P(A) stands for probability of A. Bayes' Theorem, to put it simply, is a method of determining a probability when we are aware of a given set of other possibilities.

The assumption of Naïve Bayes are no pair of variables are dependent and No attribute is irrelevant and is assumed to contribute equally to the outcome. The model is based on Bayes Theorem, but the variable X is much more complicated, see equation (10)

$$P(y|x_1, \ldots, x_n) = \frac{P(x_1|y)P(x_2|y)\ldots P(x_n|y)P(y)}{P(x_1)P(x_2)\cdots P(x_n)} \tag{10}$$

We can obtain the class using the above function with a prediction or feature. In order to calculate the posterior probability P(y|X), it is necessary first to construct a Frequency Table that compares each attribute with the target. Once the frequency tables have been transformed into likelihood tables, a Naive Bayesian equation is used to calculate the posterior probabilities for each class. Finally, the class with the highest posterior probability is selected as a result of the prediction.

3.4.3 Support vector machines (SVMs)

A supervised machine learning model called a support vector machine (SVM) uses classification techniques to solve two-group classification problems. For example, an SVM model can classify new text after giving sets of labelled training data for each category.

They offer two key advantages over more recent algorithms like neural networks: incredible speed and improved performance with fewer samples (in the thousands). As a result, the approach is excellent for text classification issues, where it is typical only to have access to a dataset with a few thousand tags on each sample. A support vector machine outputs the hyperplane—which in two dimensions is just a line—that optimally separates the tags from the input data points. Anything that falls on one side of this line will be classified as class one, and anything that falls on the other will be classified as class two. The hyperplane that maximizes the margins from both tags is found by SVM. Alternatively, to put it another way, the hyperplane (remember, it is a line in this case) with the most significant distance from the nearest element of each tag. Once it finds the best hyperplane, the algorithm is pretty much done. In this report, we use the count vectorized text data and its cluster as input. See more details about SVM in [47].

3.4.4 Performance index

After we trained our data, how should we tell the model's performance? We have accuracy, precision, recall and f1-score. Before introducing each of these meanings, I'm going to talk about four crucial concepts in machine learning:

True positives (TP): The number of data that predicted correctly and are actually positive.

False positives (FP): The number of data that predicted correctly and are actually negative.

True negatives (TN): The number of data that predicted incorrectly and are actually negative.

False negatives (FN): The number of data that predicted incorrectly and are actually positive.

Accuracy is the most commonly used score to judge a model. Unbalanced data can be problematic for the model training. It could leads to a less performed model. See function (11) for mathematical expression.

$$\frac{TP+TN}{TP+FP+TN+FN} \tag{11}$$

The precision is the percentage of positive instances from the total predicted positive instances. In other words, find out 'how much the model is right when it says it is right'. See function (12)

$$\frac{TP}{TP+FP} \tag{12}$$

Recall measures the number of positive instances out of a total number of positive instances. See Function (13).

$$\frac{TP}{TP+FN} \tag{13}$$

A F1 score measures precision and recall simultaneously. As a result, a higher F1 score is better, since it takes the contribution of both. See function (14).

$$\frac{2*precision*recall}{precision+recall} \tag{14}$$

Chapter 4 Result

4.1 Sentiment score analysis

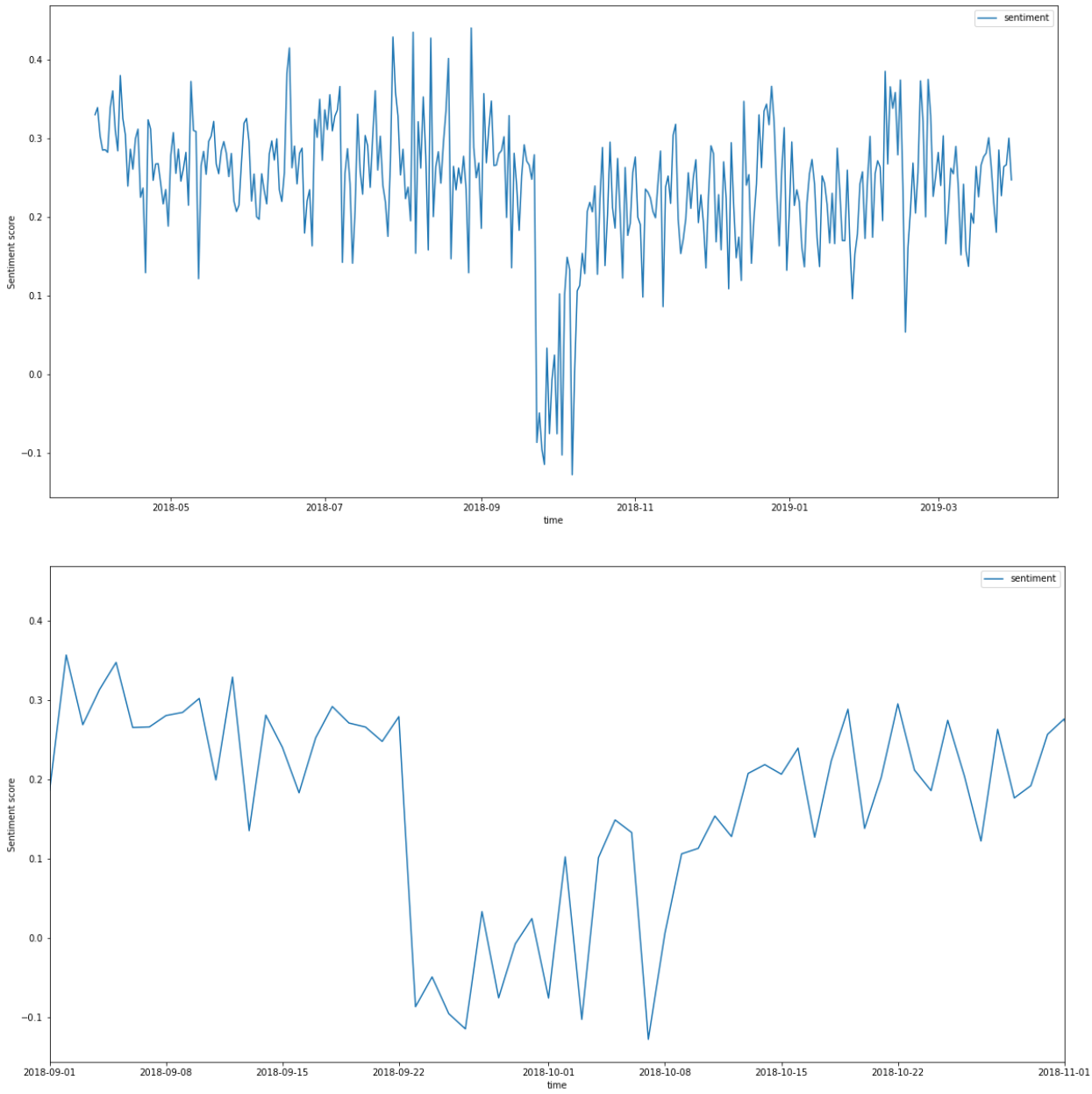

Figure 2 Average sentiment score in Pret A Manger data.

To measure the exact time of event happening in the data, we observe the change of sentiment score to do so. Figure 2 shows the change of sentiment score of Pret data with a subplot of detailed focus on the time it happened. We can see that the score is evenly distributed before huge damping around 2018-09-22. Furthermore, the event effect continues for several weeks until 2018-10-15 it returns to an understandable

stable state; however, through the first plot of figure 2. The sentiment score did not recover to the same level as before the event happened. This indicates that the event might damage the company's reputation for longer.

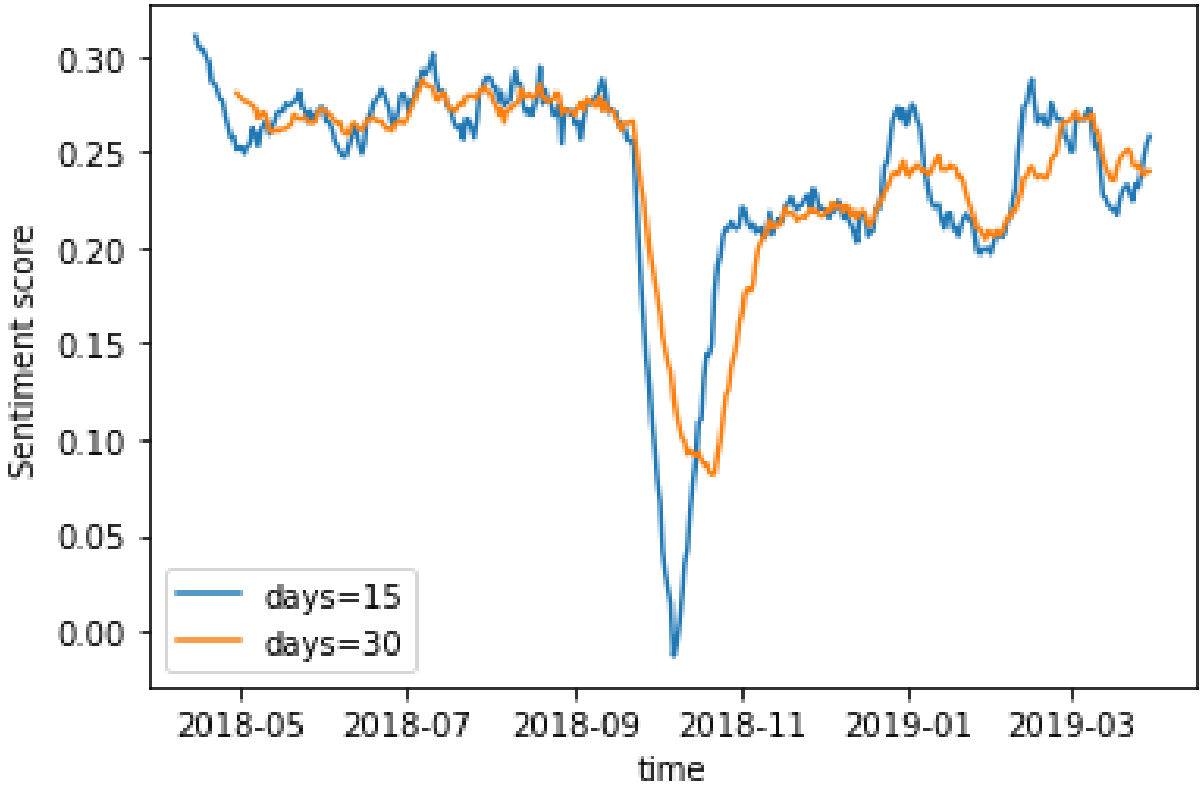

Figure 3 Moving average of Pret Sentiment Score.

A moving average is a statistical method to examine data points by averaging several subsets of the entire data set. For example, figure 3 shows two moving averages of half a month and a month. Figure 3 shows that the sentiment score has not recovered to its original value as the original average sentiment score before the event is around 0.27 and after is 0.21. This indicates that the event harmed the company's reputation. Since the sentiment score represents the general attitude of customers towards the company, also from figure 3, we can see clearly how the damping occurred when the event happened. The average sentiment score decreases from 0.27 to below 0.

After determining the time of the event, we divided the data into before and after-event datasets. Then remove all the corpus in the before dataset to the after dataset—the length of the after dataset changes from 22190 to 8975.

Let's look at another companies' sentiment score changes.

Figure 4 Average sentiment score in Ryanair data.

Figure 4 shows the change of sentiment score of Ryanair data with a subplot of detailed focus on the time it happened. We can see that the score is evenly distributed before huge damping around 2018-10-17. And surprisingly, a massive increase at the end of November 2018. The difference between Ryanair and Pret is

that for Ryanair, the event does not affect the sentiment score that much, and it quickly recovers from the dampness and even gets a considerable boost at the end of November 2018. We may assume that there is some action that Ryanair caused the quick recovery of the customer's sentiment. And we can see from Figure 5 that there is even an increase in the sentiment from an average of around 0.00 to 0.05, which indicates an increase in the reputation index of Ryanair. This is a good sign for the investor since most of the review of Ryanair is upbeat even after the shock of the event. However, we can see that compared to Pret, the average sentiment of Ryanair is lower than Pret (0.05 compared to 0.25), which means the reputation index is higher for Pret, even if the event causes a shock. At the same time, Ryanair seems to improve continuously.

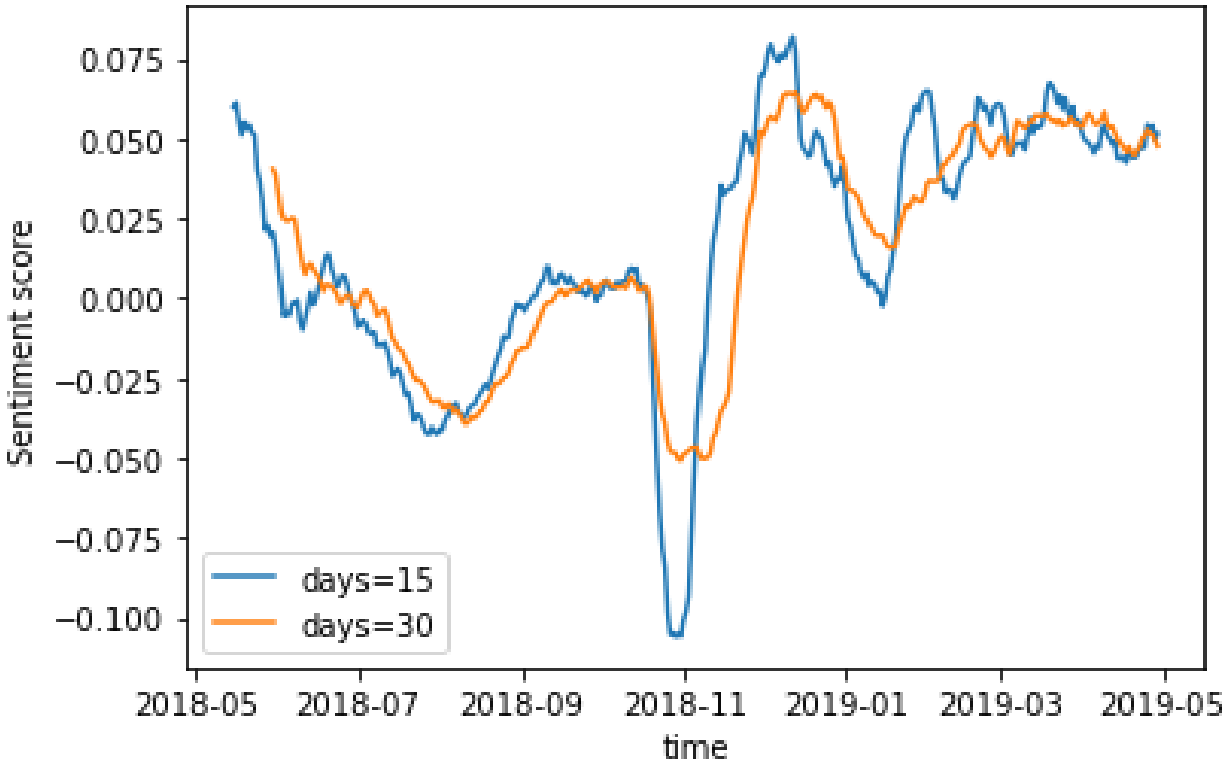

Figure 5 Moving average of Ryanair Sentiment Score

Once we figured out the exact time of the event, we can now deal with the data. We first divided the dataset into two parts-before event data and after event data. Then we deleted all the tokens that occurs in the before event data from after event data,

which will left with event-related data. And there still have some noise (event-not-related tokens) in the data we have. So that we can processed K-mean cluster to labelled the data.

4.2 K-mean Cluster

First we use Tfidfvectorizer to vectorize the corpus we have and use this as an input to fit the k-mean cluster model in sklearn library with k = 2.

We have the 20 top used tokens in the two clusters after analysis Pret data in table 2.

Cluster 0	**Cluster 1**
inquest	anj
ednanlaperous	anjindia
foodallergi	rajasthanicultur
leedsbyexampl	jaipur
neglig	indiancontemporari
loophol	citiypalac
fatal	endofseasonsal
allergyhour	thecourtyardofcolor
pretinquest	websitewwwanjkreationscom
ltltlt	anarkali
manslaught	lotu
natashaslaw	bustier
shadowban	blush
workplacebulli	asymmetr
natashaednanlaperous	therosestori
coyo	dainti
lawsuit	lehenga
jh	epitom
glyphos	sage
anaphylact	palazzo

Table 2. 20 samples of each cluster

As we can see, cluster 0 is mainly about the event since there is corpus like food allergy, fatal, lawsuit etc. We decided to choose cluster 0 as social-related data and label it in order to train our model later.

Then I applied PCA to create the plot, figure 4, to see how the clusters distributed.

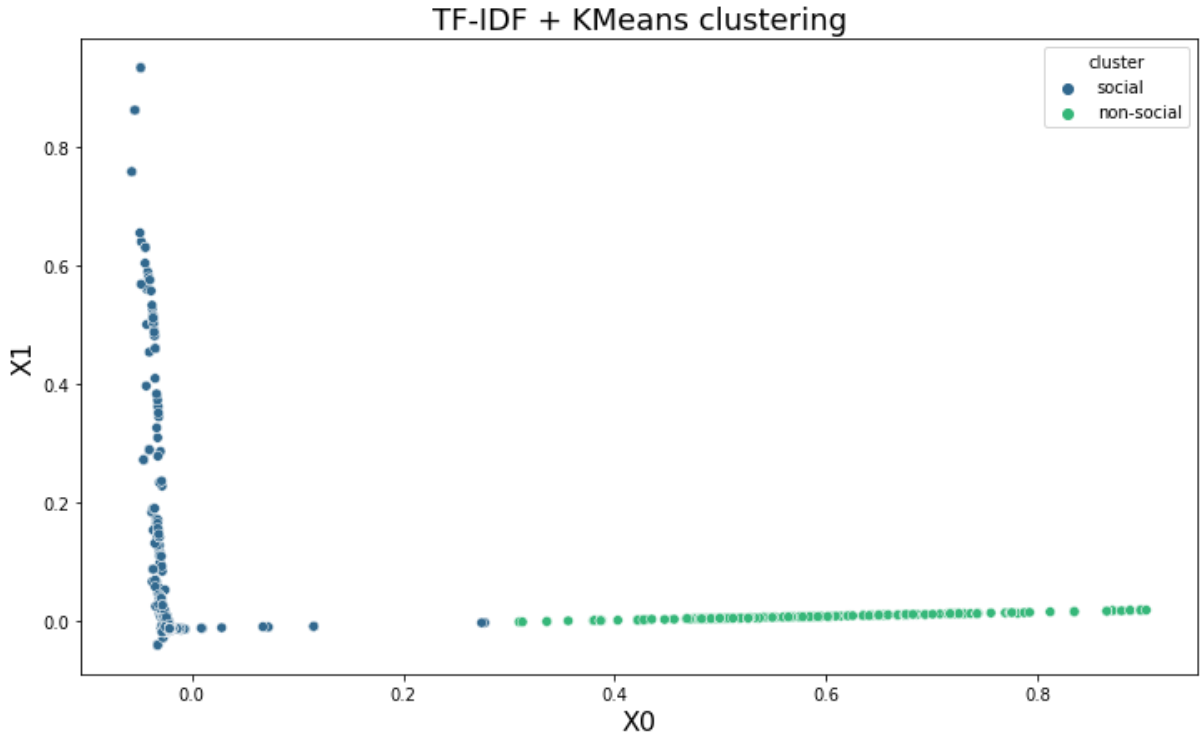

Figure 6 K-mean clustering based on TF-IDF in Pret data

In figure 6, we can see that the cluster distributed quite good since the difference between the clusters is outstanding. However, since our data is not large, the k-mean clustering algorithm should perform better if we use a large dataset. The axis X0 and X1 presented by PCA which reduced the data dimensions to 2 so that we can see clearly from a 2 dimensions figure.

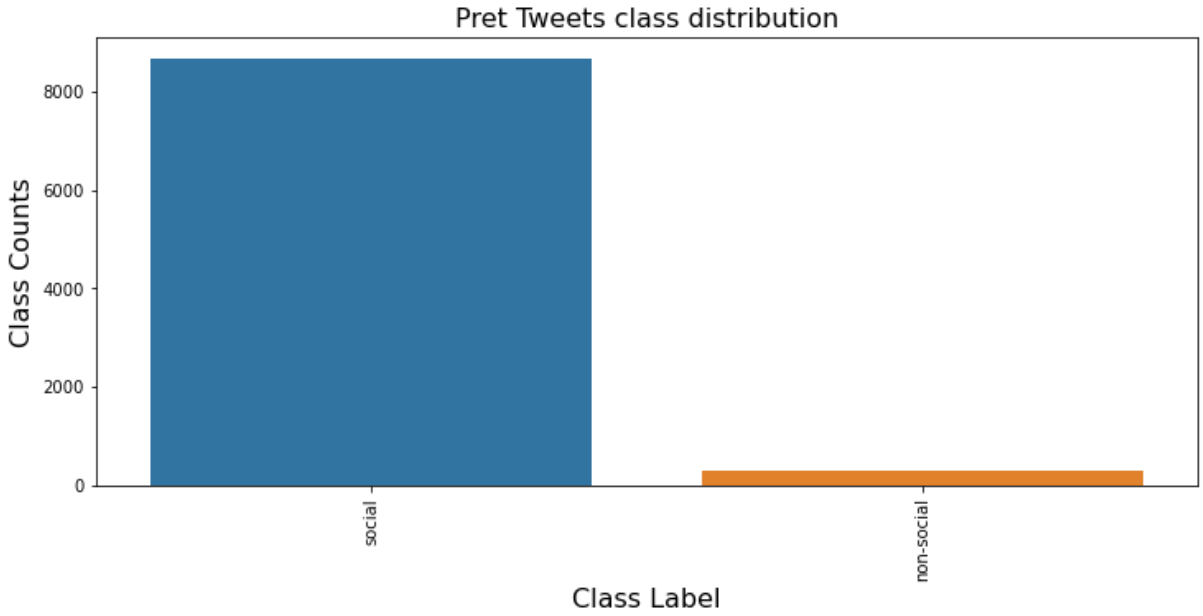

Figure 7 Pret data distribution via bar plot

In figure 7, we can see that the data is hugely uneven distributed, with 8676 data marked as social and 299 as non-social. Since the class was distributed extremely unbalanced, it may affect the model performance. Unbalanced data might be because we deleted all the background data before.

We are not processing the Ryanair data because the dataset of Ryanair is way too large, so the computer I have cannot run it in a reasonable time.

4.3 Model performance

As I mentioned in chapter 4.2 and chapter 3. We will use Logistic regression, Naive Bayes and SVM as the classification model. The reason, as stated in chapter 3, is that these three are all very suitable for binary classification problems. However, since the dataset we have is pretty small and unevenly distributed. the accuracy of the model may be misleading if the data is unbalanced. See Table 3 and Table 4.

Model	**Accuracy**
Logistic regression	96.71309192200556
Naïve bayes	96.65738161559888
SVMs	96.15598885793872

Table 3 Accuracy with each model before balancing the data

Model	**Accuracy**
Logistic regression	100.00
Naïve bayes	100.00
SVMs	100.00

Table 4 Accuracy with each model with balanced the data

Table 3 shows the accuracy of the three models using unbalanced data, we can see that logistic regression has a much similar performance as Naïve Bayes, but SVMs perform less accurate than the others. This indicated that Logistic regression and Naïve Bayes might be better for this data. As I mentioned in the last section, our data are highly imbalanced. So I tried to balance the data by randomly choosing the same amount as cluster 1 (299) of cluster 0, and sadly it will make the dataset much smaller, so the accuracy looks strange. I used Countvectorized data in all the models as x input and cluster as y input.

Table 4 is the accuracy of the three models using balanced data, as we can see that all three models performed extremely well. the detailed performances shown in table 8,9,10. And the reason of the extremely high-performance model may due to our small dataset.

The detailed performances of three models using before balanced data are shown in table 5,6,7. From table 5,6,7, we can see that all three models do poorly in predicting non-social data. Nevertheless, nearly 1.00 predict social data. The macro average represented the average value of the unweighted mean per label; the weighted average means the average value of the support-weighted mean per label. The three

models are false regarding the non-social class. From the macro average, we can see that the model acts poorly.

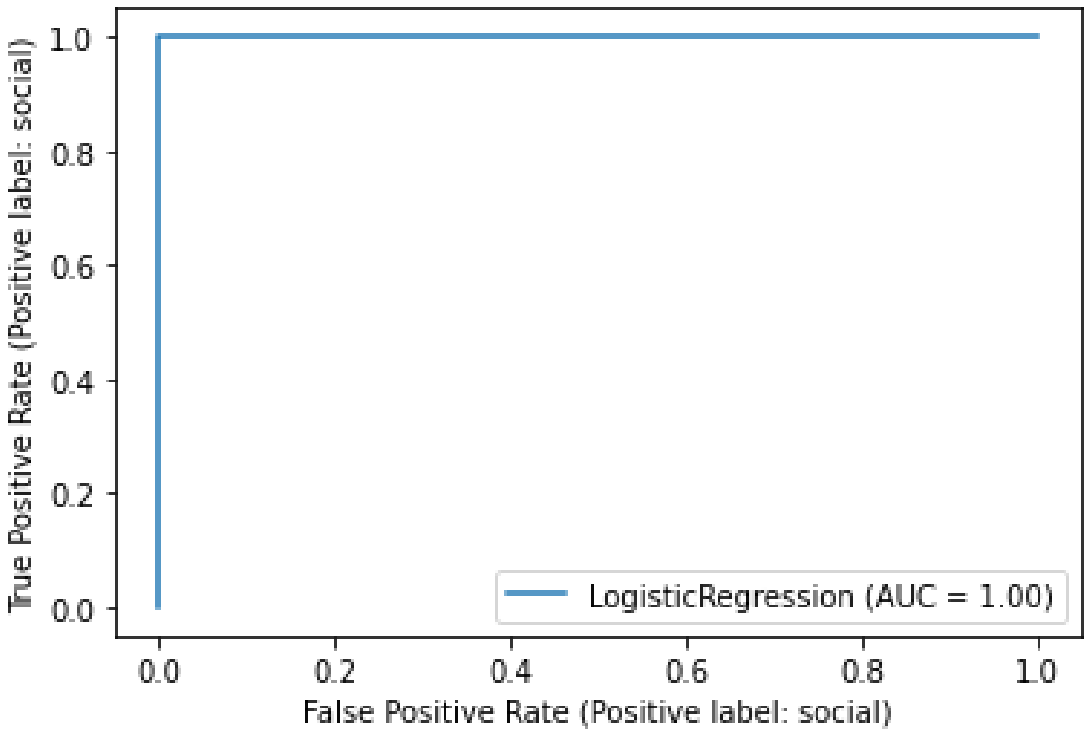

Figure 8 ROC curve of Logistic regression trained using balanced data

From Figure 8 and table 8,9,10. The performance of three models are extremely excellent. Which may highly biased since we have a small train dataset. I use the test size as 0.2. All three models predict 100% on social and non-social class, it indicates that for a small dataset, the classification model works excellent on predicting the classes.

	Precision	Recall	F1-score	support
Non-social	0.00	0.00	0.00	118
Social	0.97	1.00	0.98	3472
Accuracy			0.97	3590
Macro average	0.48	0.50	0.49	3590
Weighted average	0.94	0.97	0.95	3590

Table 5 Logistic regression classification report

	Precision	Recall	F1-score	support
Non-social	0.00	0.00	0.00	118
Social	0.97	1.00	0.98	3472
Accuracy			0.97	3590
Macro average	0.48	0.50	0.49	3590
Weighted average	0.94	0.97	0.95	3590

Table 6 Naïve Bayes classification report

	Precision	Recall	F1-score	support
Non-social	0.00	0.00	0.00	118
Social	0.97	0.99	0.98	3472
Accuracy			0.96	3590
Macro average	0.48	0.50	0.49	3590
Weighted average	0.94	0.96	0.95	3590

Table 7 SVM classification report

	Precision	Recall	F1-score	support
Non-social	1.00	1.00	1.00	68
Social	1.00	1.00	1.00	52
Accuracy			1.00	120
Macro average	1.00	1.00	1.00	120
Weighted average	1.00	1.00	1.00	120

Table 8 Logistic regression classification report using balanced data

	Precision	Recall	F1-score	support
Non-social	1.00	1.00	1.00	68
Social	1.00	1.00	1.00	52
Accuracy			1.00	120
Macro average	1.00	1.00	1.00	120
Weighted average	1.00	1.00	1.00	120

Table 9 Naïve Bayes classification report using balanced data

	Precision	Recall	F1-score	support
Non-social	1.00	1.00	1.00	68
Social	1.00	1.00	1.00	52
Accuracy			1.00	120
Macro average	1.00	1.00	1.00	120
Weighted average	1.00	1.00	1.00	120

Table 10 SVM classification report using balanced data

Chapter 5 Conclusions

5.1 Conclusions

This dissertation starts with a clear, intuitive objective explained in Chapter 1; the introduction starts with why we choose ESG to do the classification and how ESG risk affects the risk analysis. And then the explanation of why sentiment analysis is essential for operational risk management. Moreover, machine learning can be used in this area.

Then in chapter 2, the literature review, I introduced some related work on the ESG rating market, sentiment analysis, and various use of Twitter data. Starting with what is the ESG rating market, Is the market lack regulations? What will the government do to regulate the market? Next, for related sentiment analysis work, I introduced how the sentiment score is calculated by Mitic and the properties of the reputation index. Finally, I introduced several analyses with Twitter data and some research on Twitter data related to sentiment classifiers.

Furthermore, I introduced the data structure and detailed information in chapter 3 and some basic knowledge about natural language processing such as tokenization, stemming and removing stop words. Then the text embedding in order to make the text vectorized and then can be processed by the machine; I introduce the vectorized text data via TF-IDF, count vectorizer, and PCA. Moreover, the data need to be labelled by the k-mean cluster algorithm. Then the supervised machine learning so that we can produce the classification model. I introduced three models: Logistic regression, Naïve Bayes and Support vector machines. Chapter 3 mainly contains all the techniques I used for the project. I was finally shown the performance index of the model, such as accuracy, precision, recall and F1 score.

Then is the most crucial part, the result. In chapter 4, I gave the plots of sentiment analysis of Pret data and Ryanair data (figure 2-5) and the table of the result of Pret's data using the k-mean cluster (table 2). Furthermore, the visualized coordinates of the cluster (figure 6) and the distribution of the clusters (figure 7). The model

performance is stated in Tables 3 to 7. Unfortunately, the result is not that satisfying since the macro average f-1 score is 0.49 for all three models. In this case, I will explain that our research can be improved in some areas in chapter 5.2.

In chapter 1.2, I labelled six questions that this project needs to solve. I can answer them reasonable here. Sentiment analysis and ESG risk playing essential roles in recent years, especially with the ESG market, a new-born market but relatively high demand, and based on the regulations of each government, the ESG risk definitely will become the new trend in risk analysis. The classification is the way to distinguish which class the data belongs to; once the data is classified with specific signatures such as Environmental, Financial, Social etc., the risk analysis of that particular area will be introduced to the class. I believe our model can handle further data, but it needs to be improved. Currently, the model is not good enough to classify social tweets. The change of sentiment score can observe during the event time shown in figure 2-5.

In conclusion, the dissertation is an application of machine learning classification based on ESG class. It uses NLP to process the data, and our results are comparable and statistically meaningful.

5.2 Further challenges

The model we made is not perfectly worked, but the accuracy is in a good range with class social. If we have more time and better performance equipment, I believe that with large and balanced data, the model will work better. However, I did try with more extensive data: Ryanair data (nearly four times larger than Pret data). Unfortunately, my computer cannot run the tokenization and remove the stopwords part since the data is too large. More important and balanced data can improve the performance of logistic regression. Maybe we can use the validation method to improve the model's performance (for example, cross-validation)

Another direction of future work would be the way of labelling the data. In this research, I used K-mean Clustering, and I think there is a better model named BERT. It is a brand-new model introduced by Google, but it requires a much good performance computer. Also, the text cleaning part can be improved, and as you can see from table 2, there is still some unrelated corpus in the cluster. For instance, websitewwwanjkreationscom, jh, ltltlt, etc.

Finally, we could try our algorithm and test which one presents better for the sentiment score. And we could try a more binary classification model to improve the performance.

Chapter 6 Reference

[1] PRI (2021). An investor initiative in partnership with UNEP Finance Initiative and UN Global Compact PRI update Q2 2021. [online] Available at: https://harrisassoc.com/wp-content/uploads/sites/2/documents/PRI-Update-1Q21.pdf

[2] IRSG (2020), ESG Rating and ESG Data in Financial Services-A view from practitioners.

[3] P. Analytics (2018), Reputation risk in the cyber age: The impact on shareholder value.

[4] Berg, F., J. K¨olbel, and R. Rigobon (2020). Aggregate Confusion: The Divergence of ESG Ratings. SSRN.

[5] UK Government Department for Business, Energy & Industrial Strategy (2022), Mandatory climate-related financial disclosures by publicly quoted companies, large private companies and LLPs, London.

[6] Anne-Laure Foubert (2020), ESG Data Market: No Stopping Its Rise Now.

[7] KPMG (2020), Sustainable Investing: Fast-Forwarding Its Evolution.

[8] P. Krueger, Z. Sautner, and L. T. Starks (2020), The Importance of Climate Risks for Institutional Investors. The Review of Financial Studies, 33(3):1067–1111

[9] Servaes, H. and Tamayo, A. (2013). The Impact of Corporate Social Responsibility on Firm Value: The Role of Customer Awareness. Management Science, 59(5), pp.1045–1061. doi:10.1287/mnsc.1120.1630.

[10] Flammer, C. (2015). Does Corporate Social Responsibility Lead to Superior Financial Performance? A Regression Discontinuity Approach. Management Science, 61(11), pp.2549–2568. doi:10.1287/mnsc.2014.2038.

[11] LIANG, H. and RENNEBOOG, L. (2017). On the Foundations of Corporate Social Responsibility. The Journal of Finance, 72(2), pp.853–910. doi:10.1111/jofi.12487.

[12] LINS, K.V., SERVAES, H. and TAMAYO, A. (2017). Social Capital, Trust, and Firm Performance: The Value of Corporate Social Responsibility during the Financial Crisis. The Journal of Finance, 72(4), pp.1785–1824. doi:10.1111/jofi.12505.

[13] Albuquerque, R., Koskinen, Y. and Zhang, C. (2018). Corporate Social Responsibility and Firm Risk: Theory and Empirical Evidence. Management Science. doi:10.1287/mnsc.2018.3043.

[14] Chatterji, A.K., Durand, R., Levine, D.I. and Touboul, S. (2015). Do ratings of firms converge? Implications for managers, investors and strategy researchers. Strategic Management Journal, 37(8), pp.1597–1614. doi:10.1002/smj.2407.

[15] Boffo, R., and R. Patalano (2020), “ESG Investing: Practices, Progress and Challenges”, OECD Paris

[16] Securities and Exchange Commission (2022), Through the delivery of ESG-labelled securities, products, and services, trust is demonstrated in the effective integration of ESG into financial market decisions, 3235-AM96

[17] Mitic, P. (2018a). Noise Reduction in a Reputation Index. International Journal of Financial Studies, 6(1), p.19. doi:10.3390/ijfs6010019.

[18] K. Shu, A. Sliva, J. Sampson, H. Liu (2018), Understanding cyber attack behaviors with sentiment information on social media, pp. 377–388.

[19] Deloitte RiskAdvisory (2016). Reputation Matters: Developing Reputational Resilience Ahead of Your Crisis.

[20] Eccles, R., Newquist, S. and Schatz, R. (2007). Reputation and Its Risks. [online] Harvard Business Review. Available at: https://hbr.org/2007/02/reputation-and-its-risks.

[21] Mitic, P. (2017). Reputation risk contagion. The Journal of Network Theory in Finance, 3(1). doi:10.21314/jntf.2017.024.

[22] Mitic, P. (2018b). Reputation risk: Measured. International Journal of Safety and Security Engineering, 8(1), pp.171–180. doi:10.2495/safe-v8-n1-171-180.

[23] Zhao, J., Liu, K. and Xu, L. (2016). Sentiment Analysis: Mining Opinions, Sentiments, and EmotionsBing Liu(University of Illinois at Chicago) Cambridge University Press, 2015, 381 pp.; hardcover, ISBN 9781107017894, $80. Computational Linguistics, 42(3), pp.595–598. doi:10.1162/coli_r_00259.

[24] Vidya, N.A., Fanany, M.I. and Budi, I. (2015). Twitter Sentiment to Analyze Net Brand Reputation of Mobile Phone Providers. Procedia Computer Science, 72, pp.519–526. doi:10.1016/j.procs.2015.12.159.

[25] Go, A., R. Bhayani, and L. Huang (2009), Twitter sentiment classification using distant supervision. CS224N Project Report

[26] Jansen, B.J., Zhang, M., Sobel, K. and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American Society for Information Science and Technology, [online] 60(11), pp.2169–2188. doi:10.1002/asi.21149.

[27] Patodkar, V.N. and I.R, S. (2016). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. IJARCCE, 5(12), pp.320–322. doi:10.17148/ijarcce.2016.51274.

[28] Gold, N. (2020). Using Twitter Data in Research Guidance for Researchers and Ethics Reviewers. [online] Available at: https://www.ucl.ac.uk/data-protection/sites/data-protection/files/using-twitter-research-v1.0.pdf.

[29] Walters, T.N. (2008), Ongoing crisis communication: Planning, managing, and responding, wt coombs, sage publications (2007), 207 pp., paper,

[30] Purohit, H., Castillo, C., Diaz, F., Sheth, A. and Meier, P. (2014) Emergency-relief coordination on social media: Automatically matching resource requests and offers.

[31] Cameron, M.A., Power, R., Robinson, B. and Yin, J. (2012) Emergency situation awareness from twitter for crisis management. In Proceedings of the 21st International Conference on World Wide Web: 695–698.

[32] Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P.V., Fowler, J. and Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. Science Advances, [online] 2(3), p.e1500779. doi:10.1126/sciadv.1500779.

[33] meswar, M.V.Sanga., Rao, M.Nagabhushana. and Murthy, N.S. (2017). Twitter Data Analysis on Natural Disaster Management System. International Journal of Engineering Trends and Technology, 45(8), pp.394–398. doi:10.14445/22315381/ijett-v45p273.

[34] Mishra, N. and Singh, A. (2016). Use of twitter data for waste minimisation in beef supply chain. Annals of Operations Research, 270(1-2), pp.337–359. doi:10.1007/s10479-016-2303-4.

[35] Padmanayana, Varsha and Bhavya K (2021). Stock Market Prediction Using Twitter Sentiment Analysis. International Journal of Scientific Research in Science and Technology, pp.265–270. doi:10.32628/cseit217475.

[36] Doward, J. (2018). Pret allergy death: parents describe final moments with their daughter. [online] the Guardian. Available at: https://www.theguardian.com/society/2018/sep/29/pret-allergy-death-parents-demand-label-laws.

[37] Wikipedia. (2022). Ryanair racism incident. [online] Available at: https://en.wikipedia.org/wiki/Ryanair_racism_incident

[38] Post, S.M., The Washington (2017). ExxonMobil refineries are damaged in Hurricane Harvey, releasing hazardous pollutants. [online] The Texas Tribune. Available at: https://www.texastribune.org/2017/08/29/exxonmobil-refineries-are-damaged-hurricane-harvey-releasing-hazardous/

[39] Ted Baker founder and CEO Kelvin quits after misconduct allegations. (2019). Reuters. [online] 4 Mar. Available at: https://www.reuters.com/article/us-ted-baker-ceo-idUSKCN1QL0K9.

[40] Staff, R. (2021). Rio Tinto’s sacred Indigenous caves blast scandal. Reuters. [online] 3 Mar. Available at: https://www.reuters.com/article/us-australia-mining-indigenous-idUSKCN2AV0OU.

[41] Lawrence, F. (2013). Horsemeat scandal: where did the 29% horse in your Tesco burger come from? [online] the Guardian. Available at: https://www.theguardian.com/uk-news/2013/oct/22/horsemeat-scandal-guardian-investigation-public-secrecy.

[42] Wikipedia Contributors (2019). Volkswagen emissions scandal. [online] Wikipedia. Available at: https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal.

[43] K. Ela, Natural language processing (2011), IK International Pvt Ltd.

[44] FerrándezM., José Ramón Álvarez Sánchez, Félix Paz López and Toledo, J. (2013). Natural and Artificial Computation in Engineering and Medical Applications 5th International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2013, Mallorca, Spain, June 10-14, 2013. Proceedings, Part II. Berlin, Heidelberg Springer, p.92.

[45] Steven Bird, Ewan Klein, and Edward Loper (2009). Natural Language Processing with Python. O'Reilly Media Inc. https://www.nltk.org/book/

[46] Barnett, T. P. & R. Preisendorfer. (1987). "Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis". Monthly Weather Review. 115 (9): 1825.

[47] Cortes, Corinna; Vapnik, Vladimir (1995). "Support-vector networks" (PDF). Machine Learning. 20 (3): 273–297.

Project Summary

Project title: ESG classification with Tweets.

Stakeholders: Kennedys

Industrial supervisors:

Harvey Maddocks(Harvey.Maddocks@kennedysiq.com) -- Lead Data Scientist

Joe Cunningham(joe.cunningham@kennedysiq.com) -- Product Manager

College supervisor:

Fabio Caccioli(f.caccioli@ucl.ac.uk)

Team Members:

Yiqing Zhao(yiqing.zhao.21@ucl.ac.uk)

Luming Ji(luming.ji.17@ucl.ac.uk)

Ying Zhu(ucaby78@ucl.ac.uk)

Project Outline:

A 12-week summer internship at Kennedy UK is offered. The project's goal is to use Tweet data to build an ESG-based classification model. We worked as reputation advisors. The suggested approaches are supervised machine learning, NLP, and K-mean clustering.

Specific methods, including text processing, Tf-idf vectorizer, and PCA, are used to complete this project.

Mathematical expertise in probability, numerical analysis, algebra, statistical expertise in time series analysis and econometrics, and Python programming expertise in Jupyter notebooks are among the talents required.

The distribution includes the user's manual, a methodological comment, a complete description of the code, and some examples.

Project Results:

Three basic models have been contributed. Logistic regression, Naïve Bayes and SVM. Logistic regression and Naïve Bayes were identified as the best performance model.

The initial database is quite large, while I only use 2/7 of it since there is no time. Nevertheless, I'm pretty satisfied with the result, although it performed poorly in the non-social cluster.

My Contribution:

In this project, most of the work was accomplished through team collaboration, including choosing a methodology, developing initial models, calibrating the code, and selecting a calibration method. As we deal with different companies, the works are separate and not much to share. What I did in this project are:

- Learned the NLP text processing and cleaned the data.
- Outline how to clean the background data by deleting the corpus before the happening event data.
- Focus on text embedding and labelled the data with K-mean Clustering.
- Trained the model and get the result.