Imperial College London

A Graph-Based Method for Unsupervised Knowledge Discovery from Financial Texts

LREC 2022

MSCI

Joel Oksanen, Abhilash Majumder, Kumar Saunack, Francesca Toni, Arun Dhondiyal

Motivation

A financial analyst's work involves manually reviewing lengthy financial texts in order to extract relevant pieces of information. This presents a major bottleneck which could be alleviated using automated knowledge discovery and information extraction methods. At the same time, the financial services industry is heavily regulated, which means the knowledge from such systems must be accurate and explainable. This makes black-box models an unfavourable solution to this problem. Another problem is the lack of publicly available training datasets for financial knowledge discovery, which makes it difficult to use supervised learning methods in general.

We tackle these problems by proposing a novel end-to-end method for unsupervised knowledge discovery from financial texts, which can automatically generate a structured Knowledge Graph (KG) from the textual data. The KG is centred around a user-defined topic, for example sustainability. Our method can automatically analyse the resulting KG to produce numerical insights about companies in relation to the chosen topic, similar to the results of a human review of the articles.

Methodology

We focused on combining several existing methods to create a potentially useful tool, instead of developing an end-to-end model from scratch. Since the data to be extracted depends on the context, a versatile tool which can extract the relevant information just by changing the keywords, instead of having to go through the long process of model training and fine-tuning, is applicable to a larger set of scenarios. Our method takes as input a set of financial texts and a few hand-selected seed terms that help it define the KG topic. Below, we describe the four-stage process to construct the KG from these inputs. In addition, we created a novel method for the automatic analysis of the resulting KG, in relation to the chosen topic, using the semantic knowledge obtained in earlier stages.

1. Named Entity Extraction

The named entity extraction stage involves finding and extracting the named entities of interest from the input texts, which will become nodes in the KG. We focus on entities belonging to the COMPANY and PERSON classes, which we extract using an out-of-the-box Named Entity Recognition (NER) model from Stanza (Qi et al., 2020) trained on the OntoNotes dataset (Hovy et al., 2006). The dataset includes the PERSON and ORG (organisation) classes, the latter of which is a superclass of the COMPANY class. For each ORG entity, we check if it corresponds to a company identifier in our company name expansions (see box on the right) to obtain the COMPANY entities.

2. Semantic Expansion

The semantic expansion stage derives from a small hand-selected set of seed terms a large set of possible topic entities for the KG. The seed terms define the KG topic and are divided into two groups of + and -, which defines a linear topic polarity scale used in the KG analysis. The larger set of topic entities and their polarities are inferred from the seed terms utilising cosine similarities between NumberBatch (Speer et al., 2017) word embeddings.

3. Open Relation Extraction

We use the CoreNLP Open Information Extraction (OpenIE) annotator (Angeli et al., 2015) to extract open-domain (subject, relation, object) triples from the texts. For each triple, we check if the subject and object correspond to named entities or topic entities: if a match is found for both, we include the relation in our KG.

4. KG Construction

The subjects and objects of the extracted relations form the nodes of the KG, connected by the relations between them. In order to construct a useful KG, we aggregate the nodes such that in the final graph there is only one node corresponding to an entity or entity combination. The aggregation methods used for each class are detailed below.

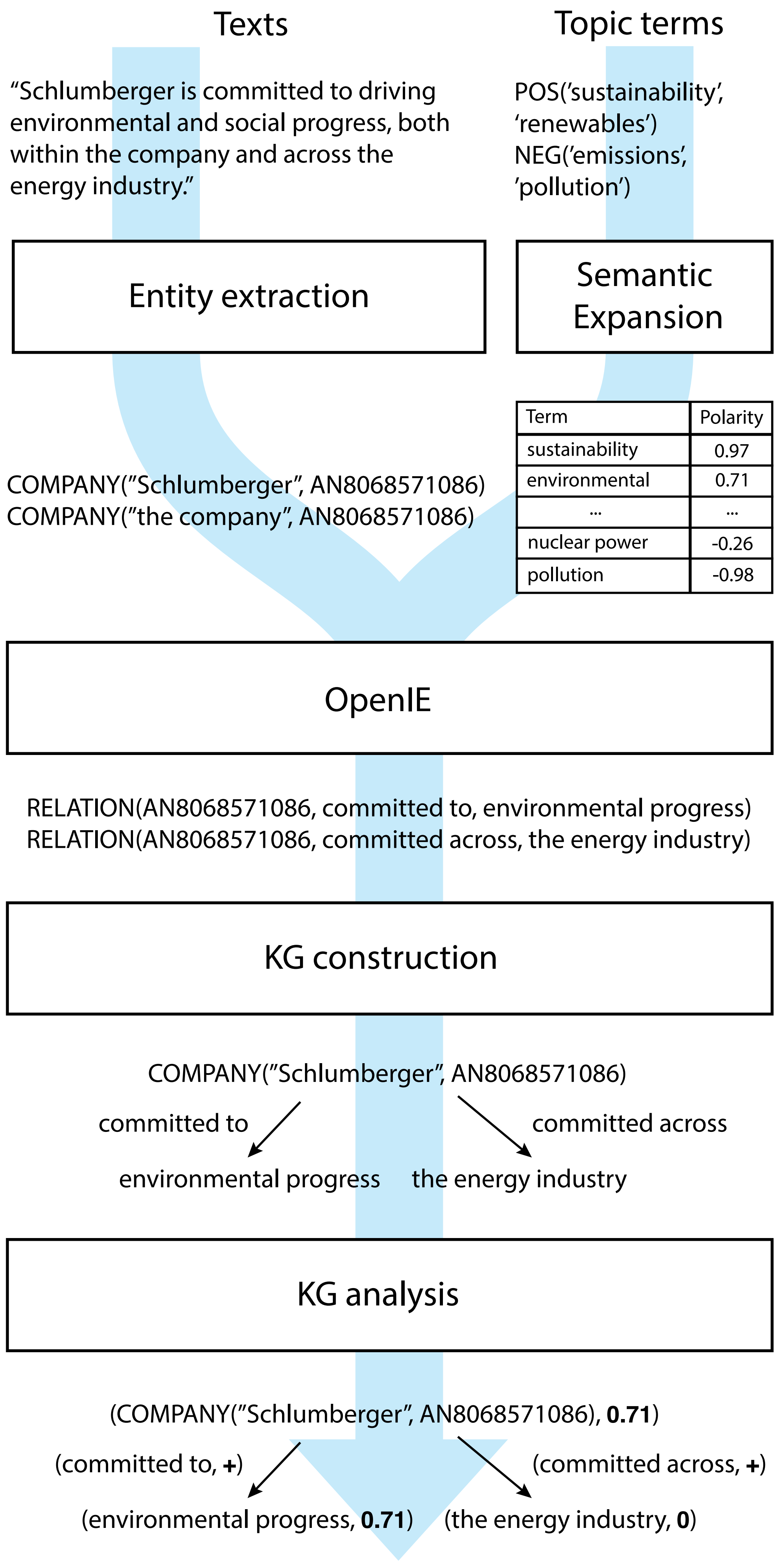
Node class	Node aggregation method
COMPANY	Company identifier
PERSON	Approximate string matching
TOPIC	Enclosed topic entities

5. KG Analysis

The knowledge graph analysis stage uses the KG to calculate a polarity figure for each company node representing the company's position with regards to the chosen topic. The figures are derived from the company nodes' relations to the topic nodes, each of which is assigned a polarity equal to the mean of the polarities of the enclosed topic entities.

Company Name Expansion

The ability to map company names to unique identifiers (ISINs) is crucial for building a concise knowledge graph with a single node for each company. While it is a trivial task to obtain an official company name for each identifier (e.g. Consolidated Edison Inc.), in natural language, companies are often referred to with several different variations of this name (e.g. Con Edison). We developed an automatic method to expand each official company name to a set of unique names by which a company might be referred to, which are used to extract company entities from texts.



Texts

"Schlumberger is committed to driving environmental and social progress, both within the company and across the energy industry."

Topic terms

POS('sustainability', 'renewables')
NEG('emissions', 'pollution')

Entity extraction

Semantic Expansion

COMPANY("Schlumberger", AN8068571086)
COMPANY("the company", AN8068571086)

Term	Polarity
sustainability	0.97
environmental	0.71
...	...
nuclear power	-0.26
pollution	-0.98

OpenIE

RELATION(AN8068571086, committed to, environmental progress)
RELATION(AN8068571086, committed across, the energy industry)

KG construction

COMPANY("Schlumberger", AN8068571086)
committed to → environmental progress
committed across → the energy industry

KG analysis

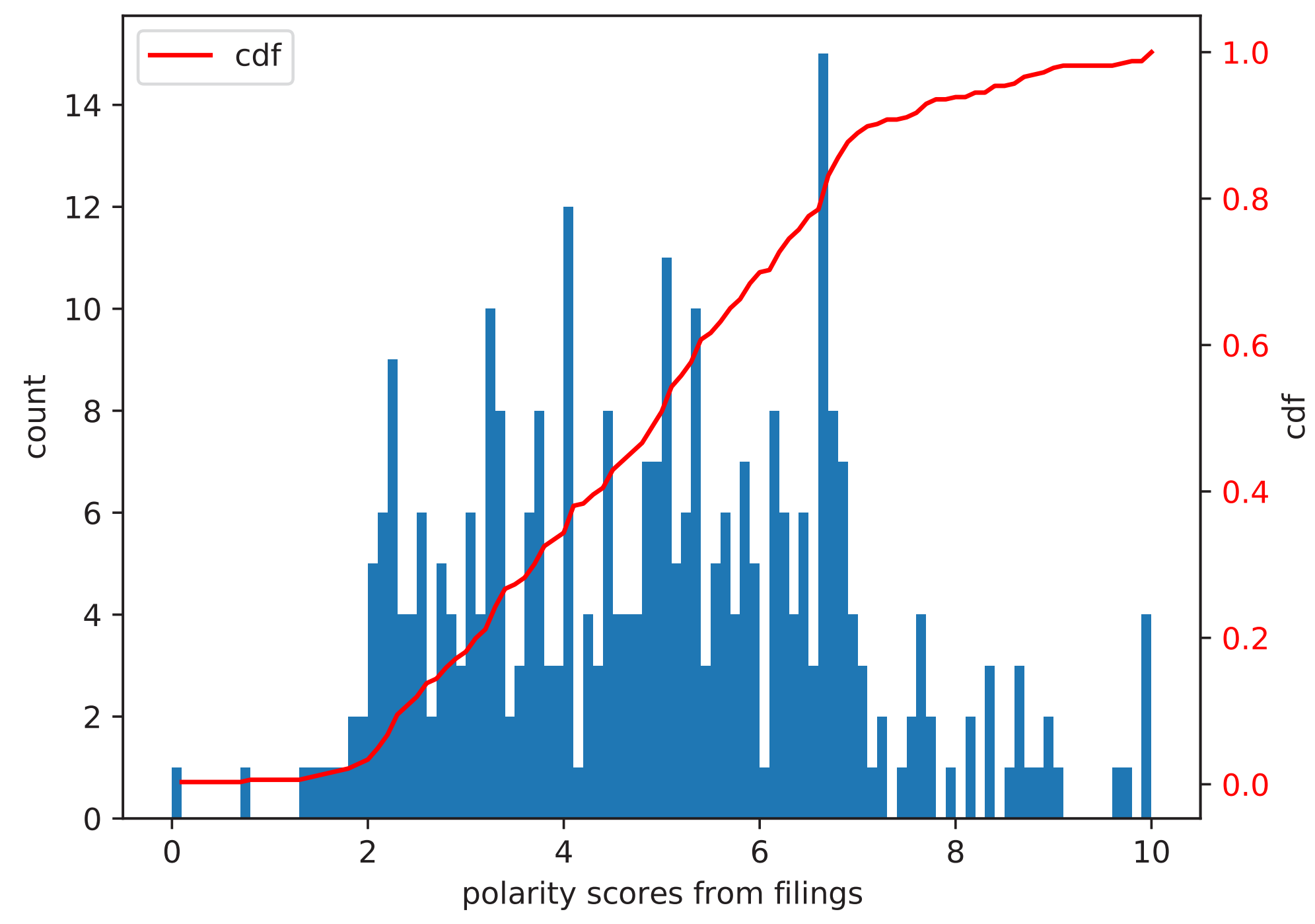
(COMPANY("Schlumberger", AN8068571086), **0.71**)
(committed to, **+**) → (environmental progress, **0.71**)
(committed across, **+**) → (the energy industry, **0**)

Case study

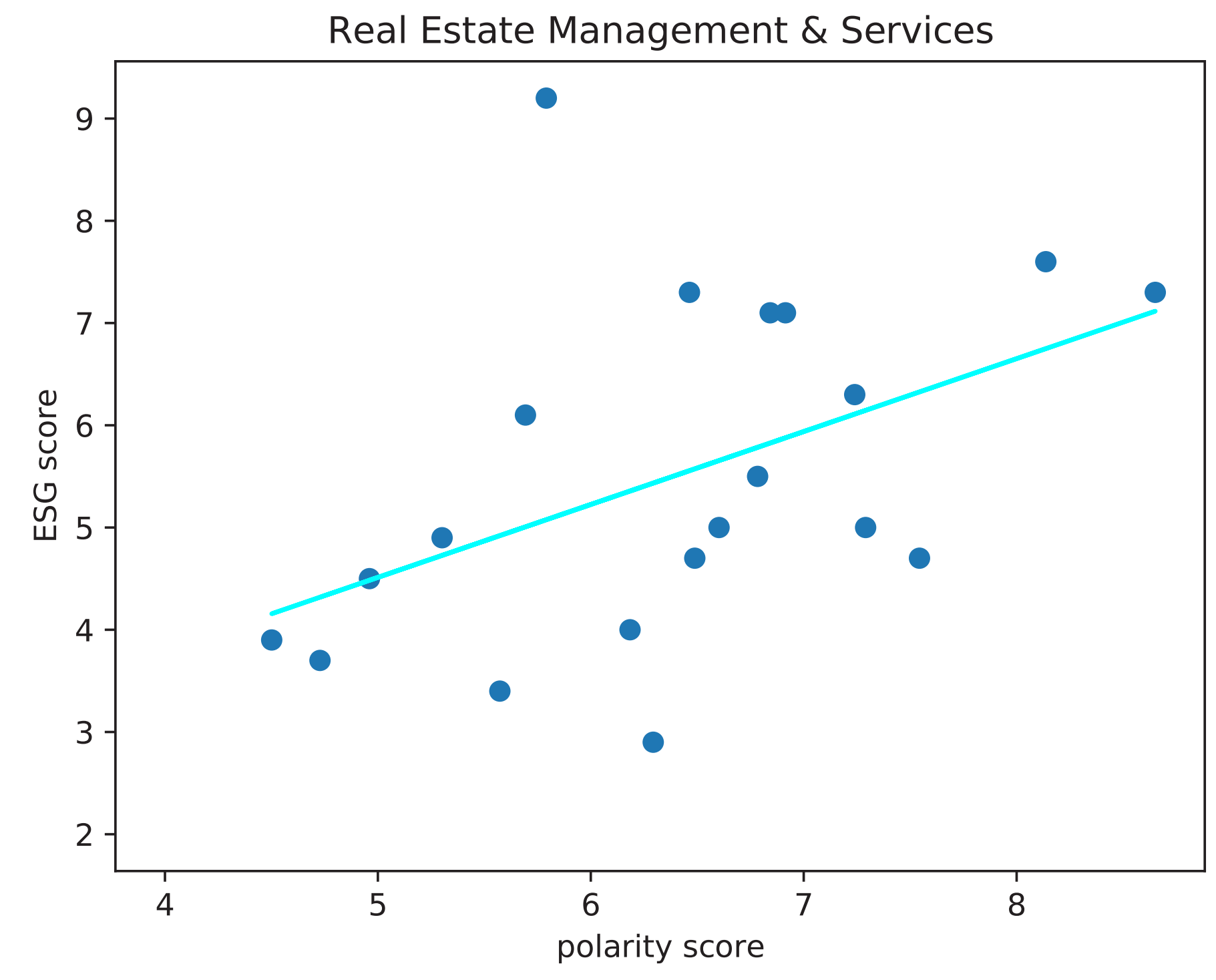
We carried out a case study to explore the usefulness of our tool for gaining insights when creating indices, and evaluated it against a standard index. We used our method to analyse the annual 10-K filings for all companies listed in the SP 500 for the previous 5 years (2016-2021). Scores were generated using the resulting graph for each company, using the following seed terms:

$$+ = \{\text{environmental, sustainability, renewables}\};$$
$$- = \{\text{emissions, pollution, fossil fuel, regulation}\}.$$

We then obtained the final scores by calculating the mean of all polarity figures over the 5 years for each company. The distribution of the scores is shown below.



Finally, we compared this output against the MSCI ESG (Environmental, Social and corporate Governance) ratings for the corresponding companies. Results from the real estate sector are presented below: the scores automatically obtained from the filings using our method follow the general trend of the manual scores assigned to the company by MSCI. However, there are variations from the trend, which are mainly due to the way that the company reports its performance in the annual reports. Some companies tend to focus more on the compliance and their achievements towards sustainability goals, while others tend to highlight this much less, and this causes fluctuations in the scoring.



Acknowledgements

This research was facilitated by the Imperial Business Partners programme at Imperial College London. The authors would like to thank Christopher J. Parker from the Imperial Business Partners programme and Akbar Hussain from MSCI for their support throughout.