Class Excercise 2

Miguel Corredera Payo (14-533-079), Patrick Stöckli (14-103-675), Tobias Hoesli (17-494-659)

2023-04-21

Introduction

In this class excercise (CE2) we are asked to use real-life scraping scenarios using the 'The Guardian' API which gives us access to over 2 million articles.

Research Question

Based on the dataset, we want to answer the following research question:

What is the relationship between articles which mention "clean energy" and articles which mention "electric car" ?

Hypothesis

Driven by climate change, renewable energy is a topic that is becoming more and more important in our society. We assume that the car industry also addresses the issue of renewable energies, as it undoubtedly has an influence on its production as well as on the marketing of electric cars.

We therefore hypothesize that: - There is a high overlap of keywords between the articles which mention "clean energy" and "electric cars".

Methods

We use the pre-made wrapper function 'gu_content' from the 'guardianapi' package to scrape articles. We tried to write a custom-made wrapper function. However, the result of the queries were non-satisfying, as we ended always ended up with the same articles, independent of the date inputs.

We first evaluate the scraped articles using an individual keyword search. This gives us a first overview of whether similar keywords are used.

We then apply a keyword network analysis to find out how the articles on the same topic are connected.

To identify the sentiment of the writers, we also apply a sentiment analysis.For that, we use the Bing, Afinn and Syuzhet lexicons which provide a polarity that sorts words into positive or negative positions with numerical values.

- The Bing lexicon has a binary categorization that simply has two values (-1 or 1).
- The Afinn lexicon rates words between -5 and 5.
- The Syuzhet lexicon has more specific values for each mood word, ranging from -1 to 1.

We scaled every result so that the range is between -1 and +1.

Our initial intention was to incorporate both the Vader and NRC sentiment analysis methodologies into our study as well. However, upon conducting preliminary tests on a small sample of 400 articles, we found that the computational time required for these methods was prohibitively long.

Results

As one can see in the results below, the expression “prime minister” was used most frequently in both topics. On the second place an again in both topics is the expression “climate change”. At first glance one can see that there are many further frequently used combinations in both topics.

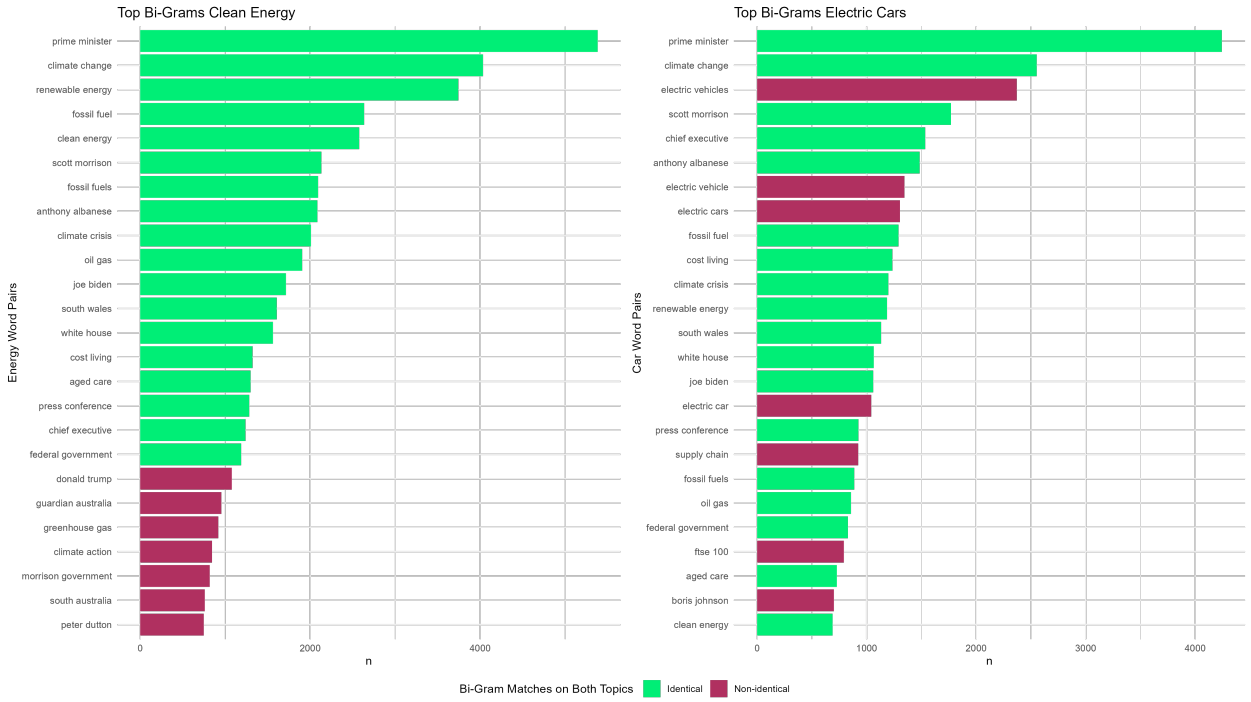

Figure 1: **Top Bi-Grams based on Topic**

On the topic of “electric car”, we are able to highlight a scattered network. The same is the case for the topic “clean energy”, but here with individual smaller networks within visible.

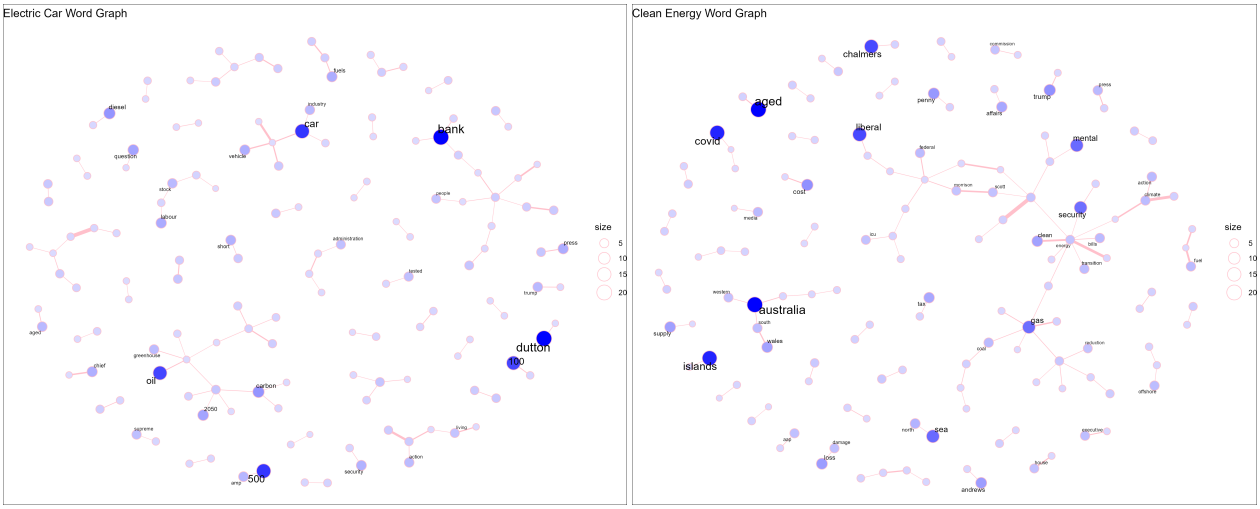

Figure 2: **Word Treegraph of Each Topic**

As one can see from the three analyses, sentiment remains fairly neutral over time for both topics. Only the Syuzhet analysis shows some rather strong changes for both topics: From mid-2022 onwards, a drop in sentiment can be seen for articles on the topic of "electric car" / "electric vehicle" and an increase in sentiment for "clean energy" / "renewable energy".

Discussion / Conclusion

We conclude that articles which mention "clean energy" and articles which mention "electric car" are very similar: They use the same keywords and have similar keyword networks. The sentiment also remains fairly neutral over time for both topics, although there are indications that this could have changed from mid-2022 onwards. It could for example be that media report more negatively about electric cars since then.

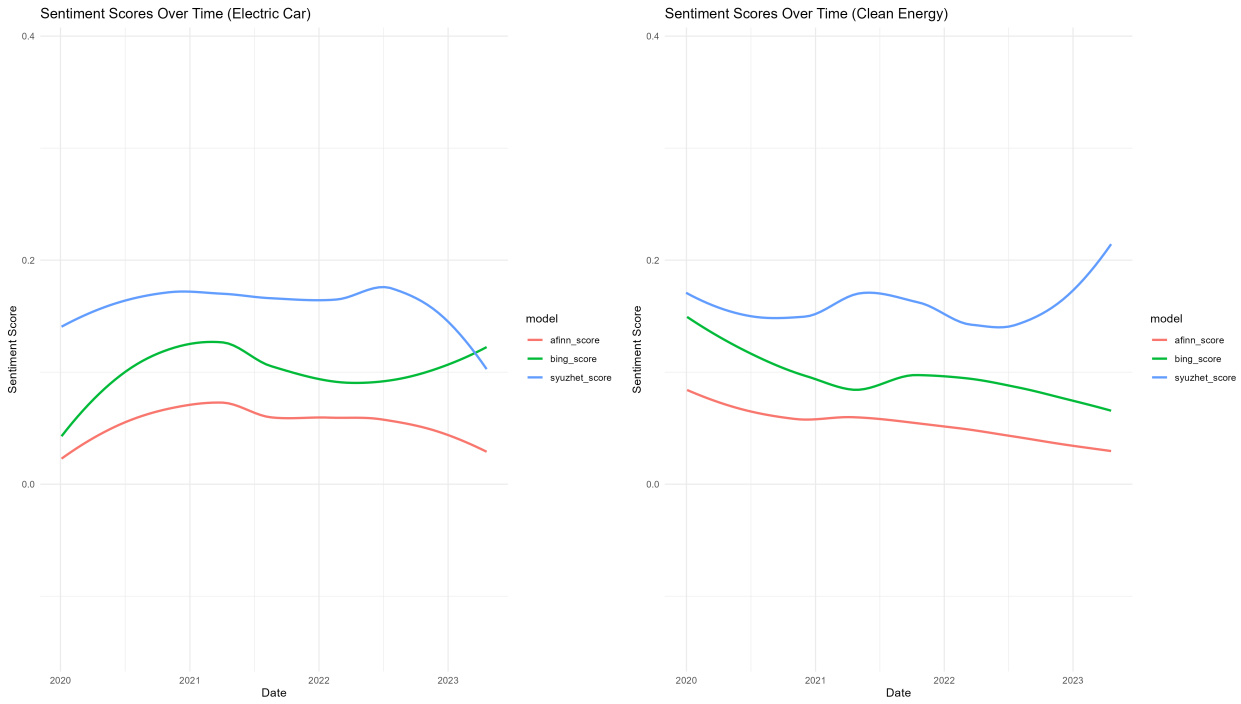

Figure 3: **Article Sentiment over Time based on Topic**