Sentiment Analysis on Robotics News

Idea

When reading news articles it seemed to me that western media reports more negatively about robotic topics than media in china does.

Initial Idea crawl social media and traditional media sites and run sentiment analysis on every text and aggregate this.

Problem 1: Crawling the Web

It's much harder then it looks. Most websites make it very difficult to extract text from them or search effectively

- E.g. Weibo prohibits copying of text directly and doesn't have a good way to search for keywords

- Reddit only gives you the last 100 entries for a given keyword

- News sites like guardian don't provide a search directly

What can we do?

There are professional tools for crawling that you can pay for.

For this project I copied a bunch of news article from websites directly

Sentiment Analysis with Textblob

Textblob is not a reliable way to do Sentiment Analysis:

E.g. This sentence is considered neutral:

Robots are revolutionizing industries and improving lives.

- Polarity=0.00, Subjectivity=0.00

We can use a large deep learning based Model instead.

Sentiment Analysis with Transformer Model

Hugginface.co provides a lot of Deep Learning / Large Language Models that you can use

I use

- "distilbert-base-uncased-finetuned-sst-2-english"
- IDEA-CCNL/Erlangshen-Roberta-110M-Sentiment

For our example sentence

Robots are revolutionizing industries and improving lives.

we get: Sentiment Label='POSITIVE', Polarity Score=0.9992

A problem with models like this is that they can only take a very limited amount of text at a time. But we can give each sentence or a couple of sentence to the model at a time and aggregate the polarity score.

Results: Sentiment Analysis with Transformer Model

Results for 5 Guardian Articles:

- Avg Polarity Score: -0.41746

Results from 14 translated Chinese Media Articles:

- Avg Polarity Score: 0.54166

Results from 14 Chinese Media Articles:
- Avg Polarity Score: 0.85693

Bias in the News

I noticed that the focus in some western news articles is often war and security related topics.

We can use spacy to find tokens that are in a list of predefined war terms

Western Media:
- 29 Mentions

Chinese Media:

- 4 Mentions

Improvements

- Using Sentiment Analysis per Sentence has the potential to misrepresent more complicated text. Here a LLM like ChatGPT could help summarize the text and then we can use Sentiment Analysis. We could also use large Models if available.

- I only a small amount of data from main stream media was used. This is not representative of the feelings of a large group of people.