Introduction

Artificial Intelligence is a fast-growing field that comes with many innovative solutions that benefit society in terms of health, transport, politics and many other fields. But like any other field of research, the world of AI is not immune to criticism and controversy. The most controversial topics related to AI often have to do with data handling, security and privacy. But even though these topics are widely seen as controversial, the way they are framed in the media can be very different. Both China and England jumped on the bandwagon when it comes to the field of AI, but how do these cultures frame this topic differently in their respective media outlets?

For this project we decided to analyze how some controversial topics within AI are framed in the media in English and Chinese. We will focus on three key terms related to controversy in our data: Privacy, Security and Data. What are the differences in terminology use between the English and Chinese authors in our dataset, and how do these differences influence the framing of the topic of AI? Read on to find out!

Data is key

First, we need to assemble a good dataset for analysis. We decided to work with one leading website focused on tech-news per language: The News Lens for Chinese, and Techradar for English. Both websites have been active for many years and have established a big brand within tech-media for themselves. They each post about a dozens of articles a day about the latest sensations in technology.

We chose to scrape 200 articles about AI in total from our two websites, giving us two datasets: one Chinese set of 100 articles, and one English set of 100 articles. Now that we have our datasets ready, let’s see what we are dealing with.

Back to basics

Before diving deeper into our analysis, it is good to lay down some basic statistics to get a feel for what is actually happening in our data. Aside from the text content of the articles, we scraped metadata on our articles

from the websites as well. Metadata is any data that provides information about one or more aspects of the data. We have compiled the metadata of our two datasets in an overview seen below.

	Publication Date	Time	Author	Title	URL	Text
1	100	100	100	100	100	100
2	91	90	27	98	100	99
3	2020-11-27	22:00:46+08:00	精選書摘	我們會被AI取代嗎？...	https://www.thenewsl...	寫到這裡，可能有人...
4	2	2	35	2	1	2

Basic Statistics: English articles

	Publication Date	Time	Author	Title	URL	Text
1	100	100	100	100	100	100
2	91	90	27	98	100	99
3	2020-11-27	22:00:46+08:00	精選書摘	我們會被AI取代嗎？...	https://www.thenewsl...	寫到這裡，可能有人...
4	2	2	35	2	1	2

Basic Statistics: Chinese articles

The tables overall look pretty similar, except for one clear difference: the amount of authors. The 100 English articles have been written by 62 different authors, whereas the 100 Chinese articles only contain 27 unique authors. When looking further into the authors of each dataset, more differences can be spotted. The English dataset contains mostly personal names of individual writers. The authors listed in the Chinese dataset however, are mostly organizations. Many of them are marketing companies or research and education centers such as TNL Marketing. Differences in our data like these are important to take into consideration, because they might influence the inferences we draw from our analyses further on.

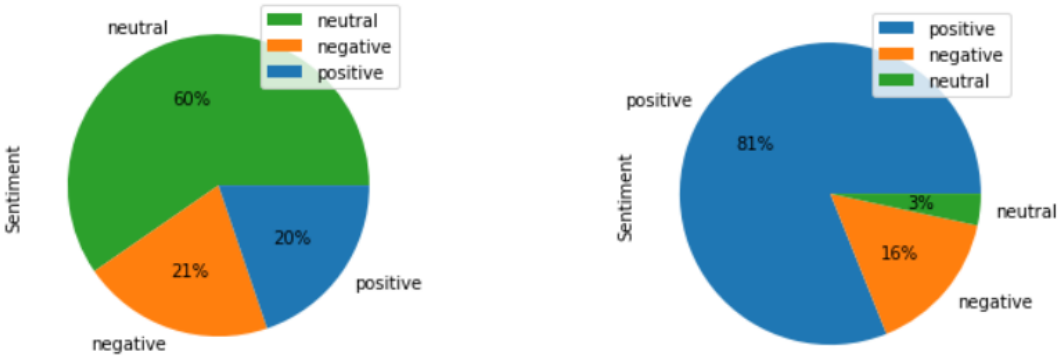
But how do they feel?

In our basic statistics, we have seen that there already some differences in how, and by who, our data was written. But what are the differences in sentiment between the datasets in our two languages?

An important part of framing is, whether the writer wants a topic to be perceived as positive, negative or neutral. Before we analyze our key terms in more depth, we decided to check the overall sentiment in our two datasets on AI. We utilized Stanza for this task. Stanza is a collection of tools made for Natural Language Processing that comes

with an option for sentiment analysis in both of our target languages: English and Chinese. We analyzed each sentence in our English and Chinese datasets, and grouped all the sentences together by their sentiment label.

After running our sentence-level sentiment analysis, we created one plot per language for comparative purposes, seen below.



Left: English sentiment plot – Right: Chinese sentiment plot

There is indeed a striking sentiment difference to be found between the two datasets. Out of all sentences in the Chinese dataset, 81% were positive sentences. This is in stark contrast to only 20% positive sentences in the English dataset. Another big difference lies in the neutral category. In the English dataset, the neutral category comes on top with 60% of sentences. For our Chinese dataset, this category is actually the smallest with only 3% of sentences.

Just from these statistics it seems that the Chinese articles have an overall goal of inspiring positive feelings towards AI. The English articles seem to focus more on neutrality and objectivity in their framing of the topic. Let's see if we can find out more information on these differences, in our term analysis.

Word relations with Gensim

Now we would like to get into our controversial key terms specifically. How are the words privacy, data and security framed in our dataset? To find out, we decided to analyze some word vector relations using the

python library Gensim. A word vector is a numerical representation of a word in a multi-dimensional space. In this space, words that are used in similar contexts will be closer together. We can use this to our advantage to spot distributional patterns in our data.

First, we preprocessed our data. For preprocessing we lowercased the text and removed all digits, stopwords and punctuations. Then, we trained two word vector models on our datasets. We trained one model on the English data and one on the Chinese data. We used a window size of 5 and set the minimum frequency to 2. Then we conducted some simple word similarity operations to see if we can spot differences in how the two languages handle our key terms.

To see how our model responds to being trained on a small dataset like ours, we first ran some similarity scores between our key terms.

English data:
(privacy , data): 0.99958515
(privacy, security): 0.99956666
(data, security): 0.99993604

Chinese data:
(隱私 privacy , 數據 data) : 0.99964285
(隱私 privacy, 安全 security): 0.9941631
(數據 data, 安全 security): 0.99457467

The similarity scores between our key terms are incredibly high. In larger datasets words usually don't get a score like this unless they are almost the same. Even though our three terms have very different meanings, since they are used in very similar contexts it seems that the model has trouble differentiating them from each other. This in combination with our small dataset gives these words higher similarity scores than is rightful in our eyes. When we run the same test on large pre-trained models, we get the following similarity scores:

English wiki news pretrained model:
(data, privacy): 0.46250108
(privacy, security): 0.577783
(data, security): 0.4066421

Chinese wiki news pretrained model:
(隱私 privacy, 數據, data): 0.88401526
(數據 data, 安全, security): 0.87624776
(隱私 privacy, 安全, security): 0.91070974

These scores look a lot more realistic. As you can see, our small dataset has made our language model's performance a lot less accurate, this is important to consider in our following analysis.

Word similarity
For our next analysis, we chose to compare lists of the top 10 most similar words for our keywords, per language. It is important to note that, since our dataset is so small, we did not get words that are actually related to the target word. However, the results can still tell us a lot about the difference in contexts in which our key terms are used in our datasets. We have compiled the following table for comparison.

	Data en	Privacy en	Security en	Data zh	Privacy zh	Security zh
0	information	smart	data	都是	終	指引
1	ai	create	nt	次	績	APP
2	one	networks	ai	包括	高	設備
3	could	make	customers	兩	名	太大
4	also	help	also	需要	想	食
5	nt	based	like	擁	第一	晚
6	make	one	microsoft	社會	機械	衝
7	use	user	re	名	率	攝
8	software	available	information	正	處	計
9	well	use	based	希望	馬	稅

Top 10 most similar words per key term

Looking at this table, the reason for the big difference in sentiment we spotted earlier becomes a lot more clear. There are some notable differences between the English and Chinese results. The three English columns of most similar terms are all quite similar, some of them containing the exact same words, and most of the words seem very business and AI related. The words also all seem quite objective in nature. They could be viewed as either positive or negative depending on the context.

The three Chinese lists of similar words seem to be quite different from each other, even though the similarity scores between our key terms in our Chinese model are just as high as in the English model trained on our data. Furthermore, the Chinese list of similar words seem to contain less words that are obviously business or AI related, and more abstract concepts such as “history” and “society”. Another difference is that some of the Chinese terms are positive by definition such as “hope”, and “good”.

Conclusion

It seems like Chinese authors employ a trend of using our controversial key terms alongside more positive words, implying that these terms aren’t all that controversial in China at all. The English authors in our dataset seem to use a more objective sentiment and a specific type of business-like terminology in relation to our key terms. Since obvious negative sentiments are not common in Western media, this objectivity paired with the lack of positive sentences points to a seriousness towards the topic. This could be due to a linguistic or cultural difference in which these two languages are expressed in news articles, or it could have something to do with the fact that most of the Chinese authors are closely linked to commercial institutions which have selling their product as a goal. For further research we would love to dive into the difference between these two, to see which one applies to our analysis, or perhaps we are dealing with a combination of both.