Presentation:

Do you have unpleasant experiences with online shopping? You got the item, but it wasn't as good as described on the website. Fake reviews may mislead you. So, is online shopping a treat or a trick? Is there any linguistic character to identify the fake reviews? I am Shiyan. In this video, I will talk about detecting fake online shopping reviews using linguistic features (ppt: topic), such as sentence structure, grammar, and word choice. Understanding the language used in reviews can help us spot the fake ones and make more informed decisions when buying items online.

(ppt: structure of the video)

This is this video’s timeline, and you can quickly navigate where you want to go.

(ppt: background and introduction)

Why?

* Pandemic makes more people choose online shopping
* Fake reviews increase with online orders, and product reviews will influence people’s purchase decisions.
* I am interested in the linguistic distinction between fake and genuine reviews

(ppt: introduce the dataset)

What?

* over 20,000 reviews generated by computers and 20,000 reviews caused by humans combined
* it has ten categories of product items
* scored on a scale from 1 to 5

(ppt: methods and results)

How

* What method we used?
* What result it tells?

(ppt: reflection)

(ppt: go to collaboratory, then introduce our program.)

(ppt: ending) Thank you for watching this video.

Presentation:

Do you have unpleasant experiences with online shopping? You got the item, but it wasn't as good as described on the website. Fake reviews may mislead you. So, is online shopping a treat or a trick? Is there any linguistic character to identify the fake reviews? I am Shiyan. In this video, I will talk about detecting fake online shopping reviews using linguistic features such as sentence structure, grammar, and word choice. ~~Understanding the language used in reviews can help us spot the fake ones and make more informed decisions when buying items online.~~

This is this video’s timeline, and you can quickly navigate where you want to go.

Why do I research this topic? First, the pandemic has made more people choose online shopping. Then, fake reviews increase with online orders, and product reviews will influence people’s purchase decisions. Lastly, I am interested in the linguistic distinction between fake and genuine reviews

I discovered that the pandemic encourages people to shop online instead of in physical stores. Then, fake reviews increased as online shopping orders increased, and product reviews influenced people's purchasing decisions. Lastly, I am interested in knowing if there is any linguistic distinction between fake and real reviews.

Next, we will talk about our dataset. As you can see, I obtained a fairly large data set about fake reviews from the OSF website. Let’s see, it contains over 20,000 reviews generated by computers and 20,000 reviews caused by humans, and it has ten categories of product items. The review text data is graded on a scale of 1 to 5. 1 is the lowest, and 5 is the highest. Furthermore, the majority of the reviews are under 50 sentences long.

Ok, it is clearly for our raw data, so let's move on to the text process. ~~To process our raw data,~~ I did four things: remove case-sensitive and punctuation; tokenize a word from text and sentences; remove stopwords; and finally, remove short words that are less than two letters long. All of these steps, when followed together, can help us create a more efficient dataset. ~~After completing these steps, we build a more formatted and efficient data structure.~~

Now that we have clean data to work with, let's start exploring deeper analyze. I chose about 2,500 reviews with less than 50 sentences, and I did a lexical diversity analysis, a frequency distribution analysis, a sentiment analysis, a bigram and an LSA analysis to find out the linguistic feature between real and fake reviews.

What are the findings?

For lexical diversity, the original has a lexicon score of 0.225, while the version made by a computer only has a lexicon score of 0.134. By comparing these two numbers, we can see that reviews written by people have a wider range of words than reviews written by computers.

Frequency distribution and word cloud

Here are the frequency distributions for both real and fake reviews. You can see that both real and fake reviews often use words like “great” and “nice”. This might mean that our evaluation texts have more positive language. The same thing can be shown by the word cloud test.

Sentiment analysis

Later, we sort our texts by rating to see if a text’s sentiment relates to its ranking. You can see that the score is higher than the sentiment, which is positive. To delve deeper into this point, we apply VADER analysis. Not surprisingly, the result matches the ratings. We also tested if article length affected polarity. They're irrelevant.

Bigram and LSA

Finally, we use Bigram and LSA to find the topics covered in the computer-generated review. We observed that computer-generated product reviews focus more on appearance, such as whether they are small or large. The same thing was found in the LSA analysis as in topic 2.

Conclusion:

Well, that is all I have for today. Let me now summarise what I talked about

Fake reviews differ in lexical diversity, frequency distribution, sentiment, and topics. This character can spot some fake online shopping reviews.

However, we don't know when the reviews will be released, and our programme has other issues, such as not comparing similarities between real and fake reviews.

That brings us to the end of the presentation. Before we wrap up this video, let's take a quick look at the programme in Colab. Thanks for watching today.

​