

Detecting Condensation

DSI 17 Capstone Project - Sarah Lim Kai Hua

Project Goals

- Identify what condensation is. Why do we want to know?
 - Not much existing research
 - Difficult to identify, so it helps the goal below
 - Make a model that can detect it
-

Example of condescension



I HAVE FOUND THE PERFECT PHRASE FOR
CONDESCENDINGLY DISMISSING ANYTHING:

HAVE YOU SEEN THE
NEW UBUNTU RELEASE?

NAH, I'M NOT REALLY
INTO POKÉMON.

- "An attitude of patronizing superiority; disdain"
- Example:
 - Post: "I don't think you know what you are talking about"
 - Reply: "Wow, that's so condescending"

Stakeholders

- Social media platforms trying to detect condescension
 - Social scientists analyzing how people converse
 - People trying to not be condescending
-



Data Source

- 5200 Posts/Reply pairs from Reddit (50/50 split)
- Labelled by people
- Corpus created by researchers in order to help research into condescension

```
@inproceedings{wang2019talkdown,  
  author = {Wang, Zijian and Potts, Christopher}  
  title = {{TalkDown}: A Corpus for Condescension Detection in Context},  
  booktitle = {Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing},  
  url = {https://www.aclweb.org/anthology/D19-1385},  
  year = {2019}  
}
```

Exploratory Data Analysis

Exploratory Data Analysis



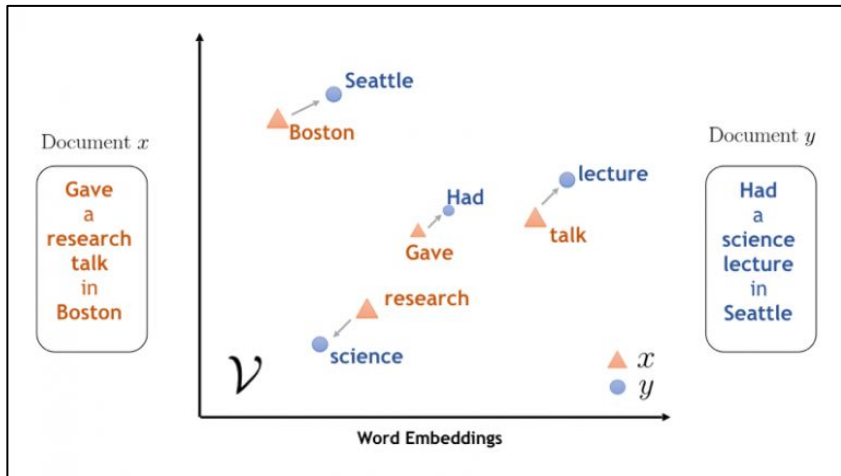
- Best way to identify - look at the reply
 - Most common responses to condescending text are the word “condescending” and swear words.
 - Least condescending words are “don’t mean”, “stay”, “took” (hard to find a pattern here)
- If the original post (not the reply) has the word ‘condescending’, it is **less** likely to be condescending
- Using reply gives a much higher level of accuracy (0.76 ROC AUC)
- Not using reply: 0.57 ROC AUC

Measuring Topic Changes

- Creators of corpus suggest that condescension causes a change in the topic of conversation, so I tried to measure this.
 - How can we determine the topic of a block of text?
 - We can use sentence embeddings
 - BERT is a way of generating these embeddings.
-

Word and Sentence Embeddings

- Word embedding: representing words so that similar words have similar representations
- Words are converted into vectors
- Sentence embedding is the same, for sentences
- BERT is able to do this (using a pre-trained model)



What is a word embedding?

How does BERT work?



Transformers - The Animation™

Look at all
other words to
get embedding

Repeat
many times

Essentially, do it
backwards to get
French*

Each layer is a
transformer

*Obviously it's a lot more complicated than this,
and we also aren't using BERT to translate



BERT

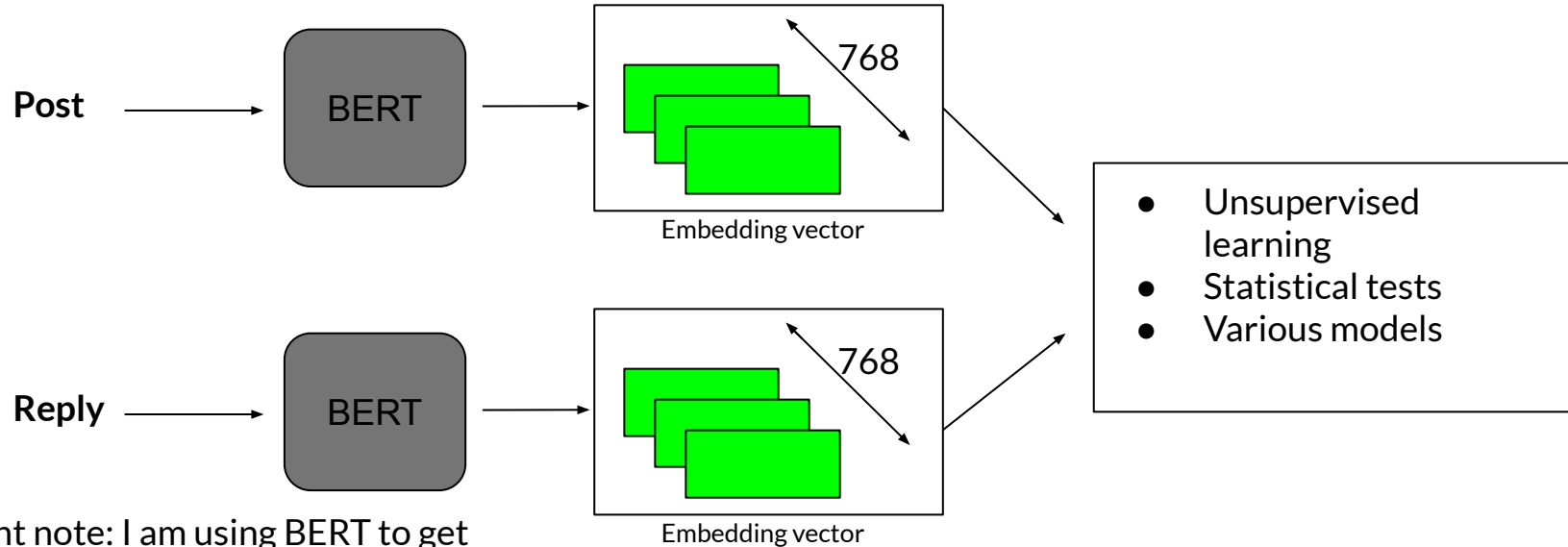
- By chaining these transformers we can get BERT
- BERT adds an additional embedding for the entire sentence
 - This is essentially the meaning of the sentence/paragraph (or at least some representation of it) in a vector form



More info on BERT

- Trained by randomly removing words from a sentence and making the model fill the blanks in
 - Don't need to have labelled data
 - Also trained on predicting which of 2 sentences is first, so it also knows sentence relationships
 - It is able to determine the context of a word (by looking at the other words)
 - Has pre-trained versions available for download
-

TL;DR - Embeddings



Important note: I am using BERT to get embeddings, so I did **not** train BERT. This is just using a pre-trained model.

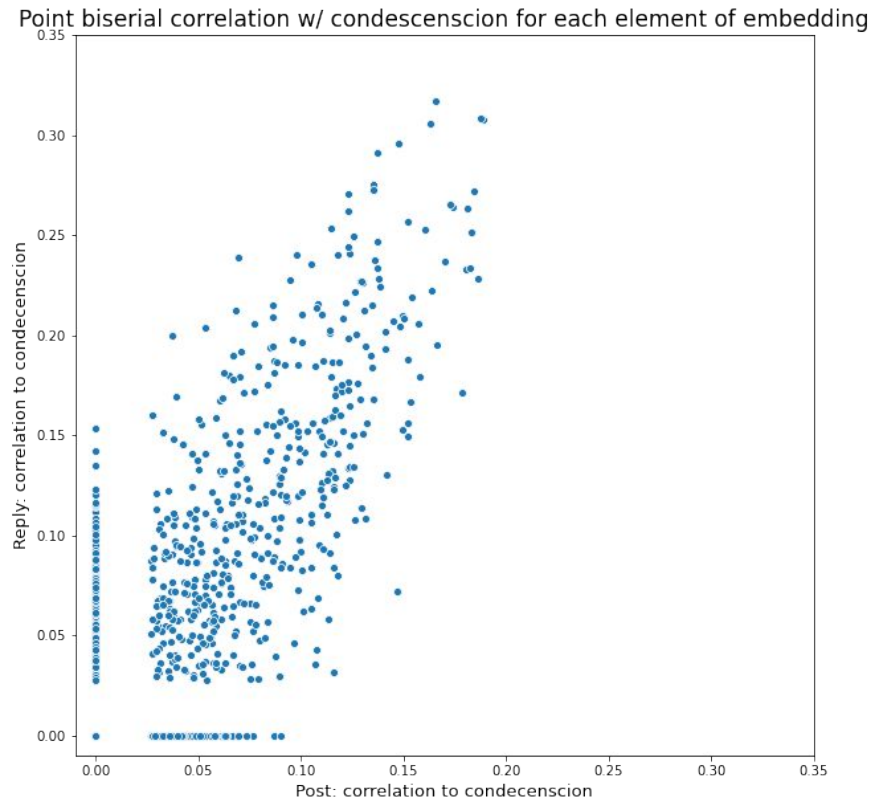
Even more EDA

—

Condescension is a cycle

- The exact meaning of each element is not easily explainable but we can still use it.
- For each of the 768 embeddings, how well does it correlate with condescension?
 - We can use a statistical test to check this
 - Plot the results of this statistical test
- Conclusion: replies to condescending posts exhibit condescending behavior

2. Are **also** highly correlated with condescension in the reply (slope is ~2 so even more condescending)

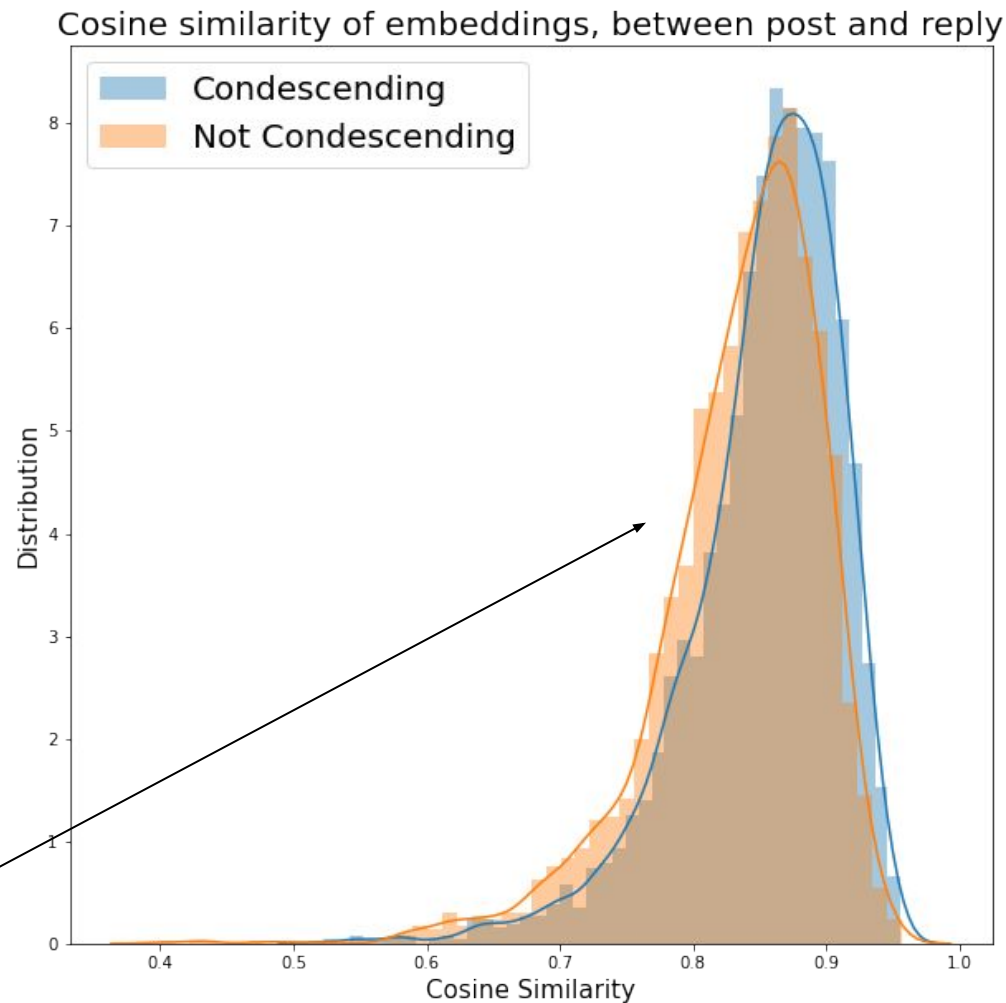


1. Embeddings highly correlated with condescension in posts

Condescension affects the topic

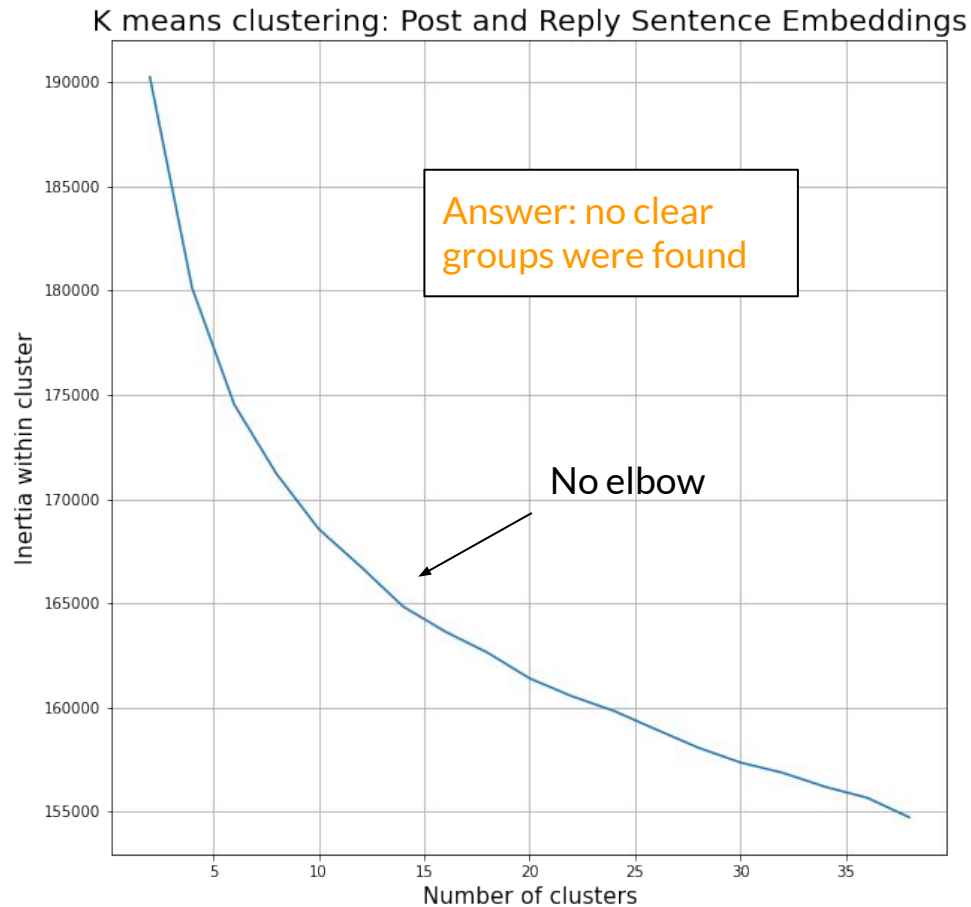
- As mentioned earlier the corpus' authors believe that condescension changes the topic of the conversation
- To plot the similarity between 2 things, we can use **cosine similarity**
- Condescending posts have more similarity

Condescending posts tend to lock in the topic



Types of Condescension

- People reply to condescension with:
 - The word “condescending”
 - Insults/swear words
- Can this reply be grouped?
- In addition, can we group condescending text into different categories?
 - Perhaps there are different ways of being condescending
- Apply unsupervised learning to the sentence embeddings



Modelling

Predicting Condensation Using Embeddings

- Basically, take these embeddings and stick them into various models (along with cosine similarity)
 - ROC AUC for models:
 - Logistic Regression: 0.78
 - Random Forest: 0.73
 - Neural net: 0.76
 - It's quite close and probably comes down to tuning hyperparameters more precisely. Logistic regression is fast so that is a good choice.
-

BERT again



- We would still like to create a model that can detect condescension directly
 - Basic models don't work very well without the reply
 - Most findings still have to do with reply, or interaction between post and reply
 - Reply is not always available, or long enough to be useful
 - Directly apply a NLP model to the **post only**
 - Use BERT again, but this time use transfer learning
 - Use '🧠 Transformers' library (yes this is actually the name)
 - This step is at the end since it loses most explainability
-

0.7
(AUC ROC)

- This only uses the post, not the reply
 - Model has somewhat high false negatives (36%)
 - Earlier models (not BERT) had ~0.57 AUC ROC
 - Still not accurate enough to reliably replace a human
-

Conclusions

- Condescending text is still difficult to classify
 - Models perform **better with access to the reply** (0.78 vs 0.7 AUC ROC)
 - Condescending speech is **replied to with more condescension**
 - People who are condescending stop the conversation from changing topic.
 - The most condescending feature is when someone replies with 'condescending'
 - However, "condescending" when used in the **post** (not reply) is very related to being **not** condescending (presumably they are self aware)
 - I had a very hard time spelling 'condescension'
-

Conclusions (for stakeholders)

- Social platform moderators: responses to condescension are also condescending, at least to our model
 - If they are taking action (e.g. issuing warnings) against condescending people, should the responses be included in this group too?
 - For social scientists: it might be worth looking at sentence embeddings
 - For longer chains of conversations
 - For other types of behavior that they might want to study (e.g. misinformation)
-

Further steps

- Corpus only consists of one post and 1 reply
 - Since condescension leads to more condescension, it may be worth analyzing entire chains of comments
 - Can also analyze how different types of replies to condescension affect the conversation (e.g. does being nice help steer conversation back on track?)
 - Requires labelled data for longer chains
 - English only has one (main) word, 'condescending'
 - Do other languages have multiple words, or different phrases to express it? Is it even a concept?
-

Q & A

(Also this is the last presentation)
