A screenshot of a social media post

Description automatically generated

Statement/Phrase Extractor

After building the keyword extractor, we thought it would be useful to provide a way to generate descriptive statements and their corresponding sentiment.

A screenshot of a cell phone

Description automatically generated

The general logic of the Statement Extractor is:

* Preprocess the text to convert “it’s” to “it is”, and “they’re” to “they are”.
* Scan through the text to find all the “is” and “are” tokens and use Spacy to extract the subtree associated with that token.
* Split the subtree into two halves: text before “is”/”are” and text after “is”/”are”
* Review all the words in the Subtree to the left of the “is” token and use POS, TAG, and DEP to look for the subject of the statement.
* Review all the words in the Subtree to the right of the “is” token and use POS, TAG, and DEP to look for the description of the statement.

When the text does not contain punctuation, the generated subtree might be pretty messy, so we have to implement stricter rules to extract coherent statements. The stricter rules also result in fewer sentences being extracted.

To evaluate the performance of the statement extractor, we made a list of all the useful “is”/”are” statements for each review video. We then generated the results for each video and calculated a score based on the equation shown below. That equation rewards points for matches and penalizes for predictions that weren’t useful and for useful statements that were missed. The maximum score is 1.

As shown in the table below, the statement extractor works pretty well with punctuated text. In this case, the score is mostly impacted by including statements that weren’t very useful. This is mainly an artifact of the reviewer being conversational and not always talking about the product.

Our model didn’t perform as well with unpunctuated review. We are losing most of our points from excluding useful statements as shown by the higher number of missed guesses. As mentioned above, this is a result of the stricter rules put in place with unpunctuated text. We were willing to omit potentially useful statements in order to better generate coherent phrases.

A screenshot of a cell phone

Description automatically generated

Per the above table, the scores for accuracy of sentence extraction **when captions include punctuation** is significantly better than when the video captions do not contain punctuation. The number of missed guesses ranged between **3%-6% for sentences with punctuation** versus **24%-39% for sentences without punctuation.**