RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Mentor Name. : John Rachlin
Subject/Category : Computer and Information Sciences
Class : Advanced Programming with Data
Group # : 58

Northeastern University

Biases and Language Differences in Research Papers on Nicotine & Tobacco Safety

Eddie Lowney, Nick Usich; John Rachlin

BACKGROUND

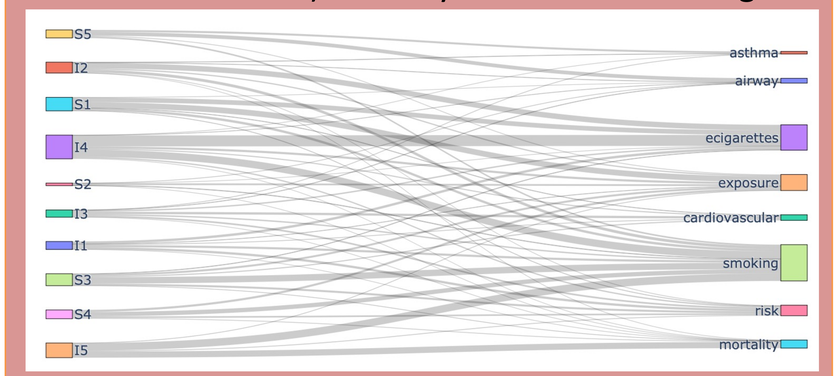
Tobacco companies' infamous disinformation campaign to muddy research on tobacco health effects has slowed in recent decades. However, these same companies still fund research on tobacco and nicotine health effects. We sought to quantify the influence this funding has on sentiments and conclusions of research papers.

METHODS

To investigate how much funding from tobacco companies influences research, we collected 5 papers which did receive funding and 5 which did not. For our framework, we created a library for the processing of any text and creation of NLP visualizations which we implemented to analyze the 10 papers.

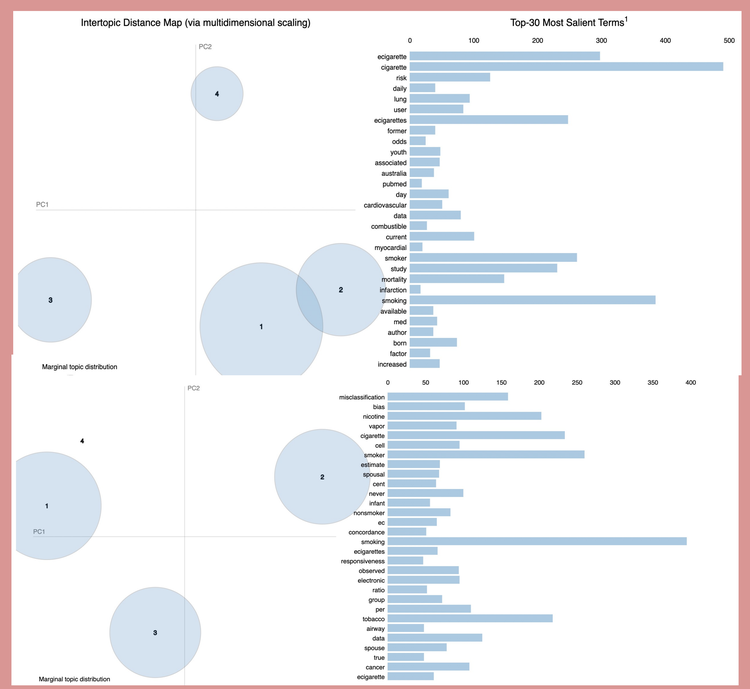
FINDINGS

We first implemented a couple of surface level techniques showing the language and sentiments being used in the papers (Sankey, VADER sentiment scores). These yielded minimal insights.

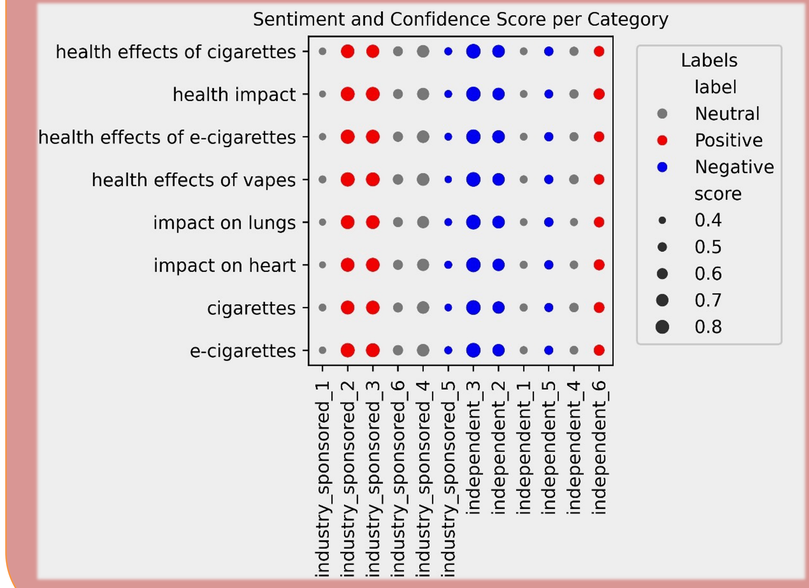



```
[('I1', -0.0008894002789400279), ('I2', 0.003557744209466264),
 ('I3', 0.010175598273832876), ('I4', 0.007366991587661904),
 ('I5', 0.0011550200803212853), ('S1', 0.005872794117647059),
 ('S2', 0.011788667937891583), ('S3', 0.007353613945578232),
 ('S4', 0.015465841406530319), ('S5', 0.0314248297537978)]
```

LDA distilled each document into 4 topics. The topics overlapped significantly between groups.



We then used the ChatGPT API to identify key portions of the texts and used an ASBA model to generate sentiment scores. ASBA analyzes the sentiment on certain topics within the full text.



CONCLUSION

The differences in sentiment between independent and industry funded research seen in the visuals, especially in the dot plot, suggest that industry money may influence scientific research on tobacco health effects. If expanding this project, we would seek a larger sample of papers, as with only 10 papers, our findings could easily be due to chance. We would also analyze the papers' exact findings to determine whether our polarity scores match with the actual research conclusion. This would allow us to assess the accuracy of our polarity scores.

Author Contributions: Eddie: Loaded and cleaned data, generated the Sankey diagram, conducted VADER sentiment analysis, and worked on the ASBA model.
Nick: Researched papers for analysis, used OpenAI API to locate key sections, worked on ASBA model and dot plot, and used LDA model.