DATA 25900

FINAL PRESENTATION

DEAN CASEY & MEGHANE SAIDENBERG-CORISTON

Research Question

Do LLMs trained on different bodies of text generate headlines for news stories with inherently different political biases?

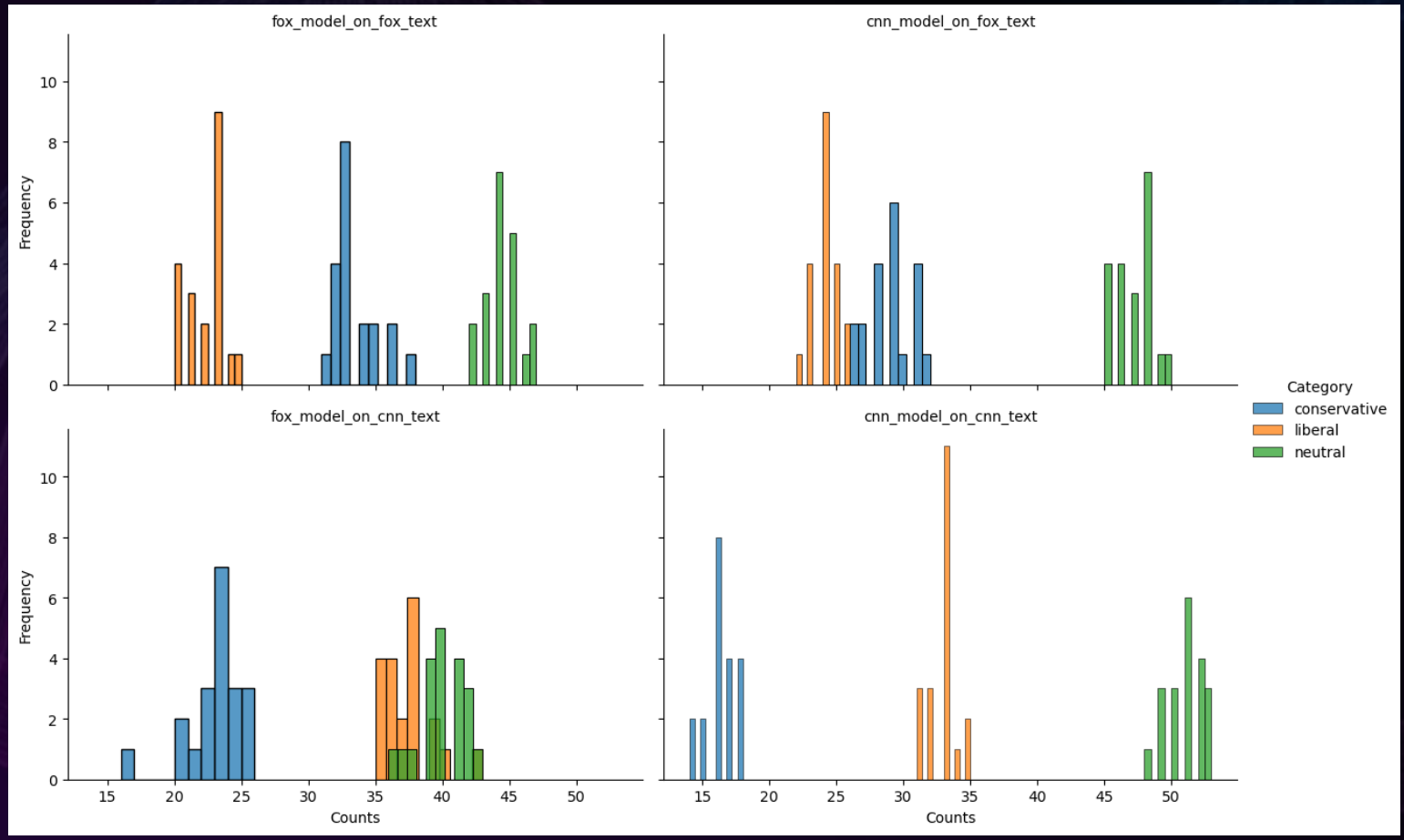
Methodology

Web scraping Fox News and CNN politics articles

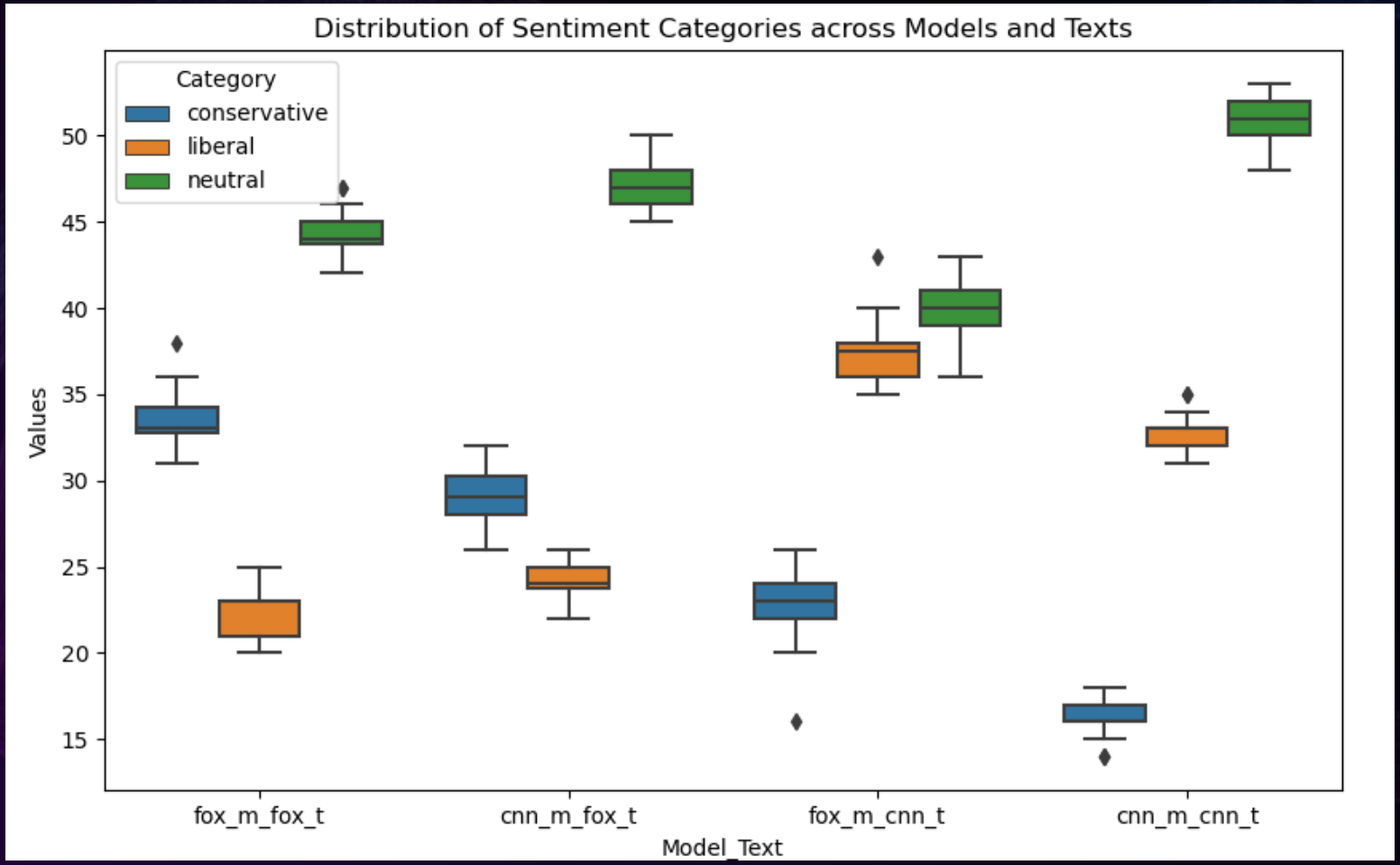
Training two LLMs on respective bodies of data

Generating test headlines on fine-tuned LLMs

ANALYSIS I: GENERATIVE MODEL



ANALYSIS I: GENERATIVE MODEL



ANALYSIS II: DISCRIMINATIVE MODELS

	Multinomial Naïve Bayes	Naïve Bayes	Logistic Regression	Linear Support Vector Machine
All test data (Fox model with CNN stories, Fox model with Fox stories, CNN model with Fox Stories, CNN model with CNN stories)	mean of simulations: 0.554 standard deviation: 0.04105151973712146 t-statistic: 1.315420241340535 p-value: 0.09494495251892054	mean of simulations: 0.549125 standard deviation: 0.03400158752978241 t-statistic: 1.4447854811770415 p-value: 0.07504941355509742	mean of simulations: 0.571125 standard deviation: 0.041503278135101525 t-statistic: 1.7137200528708552 p-value: 0.04407258968463945	mean of simulations: 0.57075 standard deviation: 0.03706481491995482 t-statistic: 1.9088183807956869 p-value: 0.028865870584248277
Heterogenous model + text Combination (CNN model with Fox stories and Fox stories with CNN stories)	mean of simulations: 0.5295000000000001 standard deviation: 0.07065530211648463 t-statistic: 0.4175199753780038 p-value: 0.3383752505190907	mean of simulations: 0.5259999999999999 standard deviation: 0.06935125059864428 t-statistic: 0.37490311675083443 p-value: 0.3540668877230435	mean of simulations: 0.5315000000000001 standard deviation: 0.06312765645975629 t-statistic: 0.49898890227425513 p-value: 0.3091705672755751	mean of simulations: 0.5262499999999999 standard deviation: 0.06320311573319656 t-statistic: 0.41532762579001203 p-value: 0.33917590031648026
Homogenous model + text combinations (CNN model with CNN stories and Fox model with Fox stories)	mean of simulations: 0.606 standard deviation: 0.06186495556672394 t-statistic: 1.713409458213779 p-value: 0.044101193919563575	mean of simulations: 0.5742499999999999 standard deviation: 0.06699304577206638 t-statistic: 1.1083239930995856 p-value: 0.13453319293883492	mean of simulations: 0.61675 standard deviation: 0.072217317238256 t-statistic: 1.6166482564676816 p-value: 0.053773171289072774	mean of simulations: 0.62125 standard deviation: 0.06594064901044873 t-statistic: 1.8387747439487152 p-value: 0.033723063262548036

EVALUATION

ANSWER TO QUESTION: ...MAYBE?!?!

PITFALLS AVOIDED

- Rigorous data cleaning process
- Data corpus matching
- Open declaration of result limitations – transparency in discussion of result veracity

LIMITATIONS

- Web scraping methodology inconsistent along CNN and Fox News
- Limited data used to train generative and discriminative models
- Crude metrics for success

Thank you!!

Thank you for listening!! ☺