CS 2731 / ISSP 2230
Introduction to Natural Language Processing

Session 16: Project proposal presentations

March 4, 2024

School of Computing and Information

Schedule

1. Shiyuan, Yingda
2. Hongtao, Chonghao, Sean, Bo-Chen
3. Werner, Yuelyu, East, Anfeng
4. Yuning, Na, Ken, Yuelong
5. Sai, Deyasini, Shiva, Aparna
6. Purva, Fatemeh, Ayush, Shayan
7. Kartik, Kasvitha, Brian, Owen
8. Nick, Arushi, Trung
9. Noah, Annanya, Jayden, Xiaoyan

Instructions

- Plan for **5 min presentations** not including Q&A
- Cover at least these key points
  - Project motivation (what is the value of this work?)
  - Super briefly, what 1-2 other related papers have done
  - What data you are planning to use
  - What approach/methods will you be taking
  - Evaluation of your approach (or dataset, if it's a dataset contribution)
- Put your slides in this presentation after your project name slide by **class session, 3:00pm on Mon Mar 4**

1. Shiyuan, Yingda

Motivation

- A new mechanism called the attention constraint mechanism through auxiliary attention.
- This method is designed to enhance the performance of translation tasks by making the distribution of attention weights more precise and focused, addressing the issue of fuzziness commonly found in existing models.
- Our approach involves adding an auxiliary attention layer and applying a mathematical constraint to refine the focus of the model, leading to better translation and summarization outcomes without significantly increasing the complexity of the models.

Related papers

- *[2] B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang, "Modeling localness for self-attention networks," in EMNLP, 2018, pp. 4449–4458.*
- *[3] Z. Fan, Y. Gong, D. Liu, Z. Wei, S. Wang, J. Jiao, N. Duan, R. Zhang, and X.-J. Huang, "Mask attention networks: Rethinking and strengthen transformer," in NAACL, 2021, pp. 1692–1701.*

We extend research efforts such as those presented in papers [2] and [3], which aimed to directly improve attention distribution through new mathematical expressions. Specifically, [2] introduced Gaussian bias for self-attention, and [3] proposed new mask matrices to enforce localness modeling ability. Our work diverges by moving away from fixed mathematical expressions to introduce a constraint relationship that adapts during training, thus altering the attention distribution more dynamically.

Dataset

- IWSLT Zh-En dataset
- contains about 170,000 training sentence pairs, 7,000 valid pairs, and 7,000 test pairs.
- For the WMT En-de tasks, the data sizes are 4.5 million.
- For WMT En→De, sentences are encoded by a joint source and target vocabulary of 37,000 tokens.

Approach & Evaluation

- A novel attention constraint mechanism, integrated with an auxiliary attention mechanism
- Evaluation:
  - BLEU Scores: For neural machine translation tasks, we are using BLEU scores as the primary metric to assess the quality of our translations
  - Ablation Study: We conduct an ablation study to understand the impact of each component of our proposed approach on the overall performance.
  - Parameter Increase: We also evaluate the efficiency of our approach in terms of the increase in the number of parameters compared to the base models.

2. Hongtao, Chonghao, Sean, Bo-Chen

Hate speech detection with Contrastive Learning

Sean Linton, Bo-Chen Kuo, Chonghao Qiu, Hongtao Wang

University of Pittsburgh.

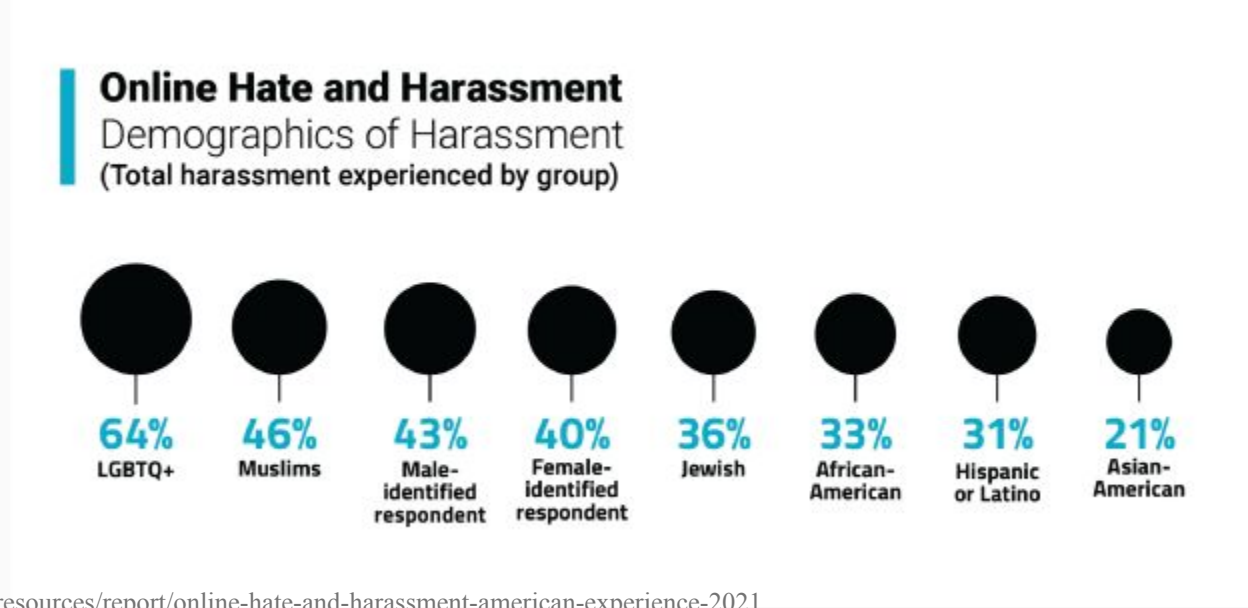
Content

1. Motivation
2. Related works
3. Data
4. Methods
5. Evaluation

Motivation

Source: https://www.un.org/en/hate-speech/

Motivation



Source: https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021

Related Works

- A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media
- Generalizable Implicit Hate Speech Detection using Contrastive Learning
  - Details the method for fine-tuning BERT using contrastive learning

Data

- Implicit Hate Corpus (IHC)
  - not hate, implicit hate, explicit hate.
- Social Bias Inference Corpus (SBIC)
  - Social Bias Inference Corpus
- DynaHate
  - hate, not hate

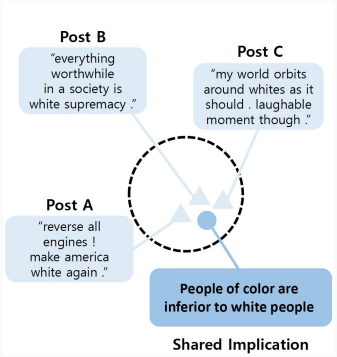
Methods

BERT fine-tuned with contrastive learning:

- Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (https://aclanthology.org/D19-1410)

Contrastive Learning:

- SimCLR(T. Chen et al.)
- Biased Toxic keywords and implicit hate speech



T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations,"

Evaluation

- Macro-averaged F1 score as the main performance metric
  - Balance minimizing false positives and false negatives
- Perform several cross-evaluation experiments
  - Using different combinations of BERT models, training datasets, loss functions, and hyperparameters
  - Measure F1 score for these combinations and compare results

End

Thank you! Questions?

3. Werner, Yuelyu, East, Anfeng

Error Detection in Medical Notes

By Pengyu Chen, Werner Hager, Yuelyu Ji, and Anfeng Peng

Motivation

Errors in clinical notes and misdiagnoses can cause a variety of issues, such as wasted medicine, delayed or harmful treatment, and could even result in major harm or death.

General purpose LLMs currently struggle with identifying these forms of specialized errors in text.

Explore how to implement LLMs using fact verification and commonsense reasoning for applications that require professional knowledge.

Dataset and Evaluation

Dataset

- **Primary Dataset**: MEDIQA-CORR: Medical Error Detection & Correction competition, which contains 2761 labeled cases.
- **Secondary Dataset**: BioASQ, supplying external knowledge

Metrics:

- F1 score
- Recall

Benchmarks

- BERT
- Llama

Dataset example:

Statement: Blood cultures are sent to the laboratory. Intravenous antibiotic therapy is started. Transesophageal echocardiography shows a **large, oscillating vegetation** attached to the tricuspid valve. Causal organism is Staphylococcus epidermidis. There are multiple small vegetations attached to tips of the tricuspid valve leaflets. There is moderate tricuspid regurgitation. The left side of the heart and the ejection fraction are normal.

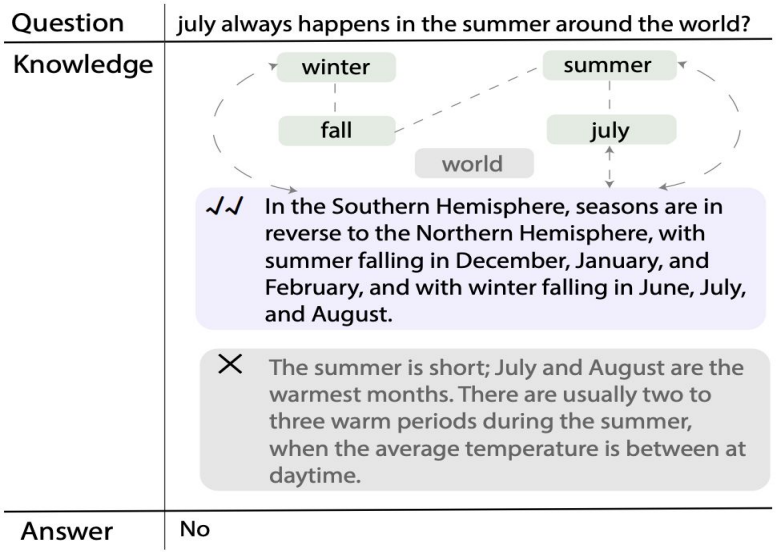
External knowledge:

1. Staphylococcus aureus is a much more common pathogen, especially when large, oscillating vegetations are involved.
2. Staphylococcus epidermidis. usually associated with medical device-related infections and the infections it causes are **usually milder.**

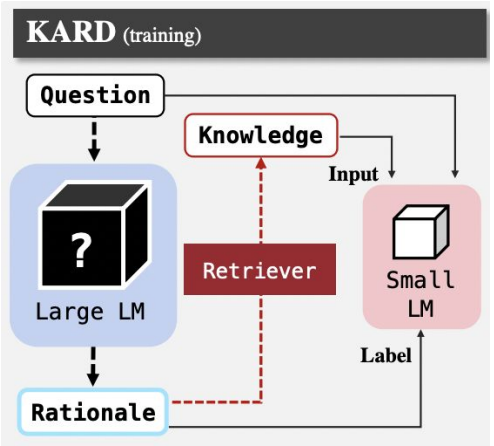
LLM generate rationale: “Virulence Factors: Staphylococcus aureus is generally more virulent than Staphylococcus epidermidis and is more often associated with the formation of large, oscillating vegetations on native valves.”

Key Papers

Decker: Double check with heterogeneous knowledge for commonsense fact verification



Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks

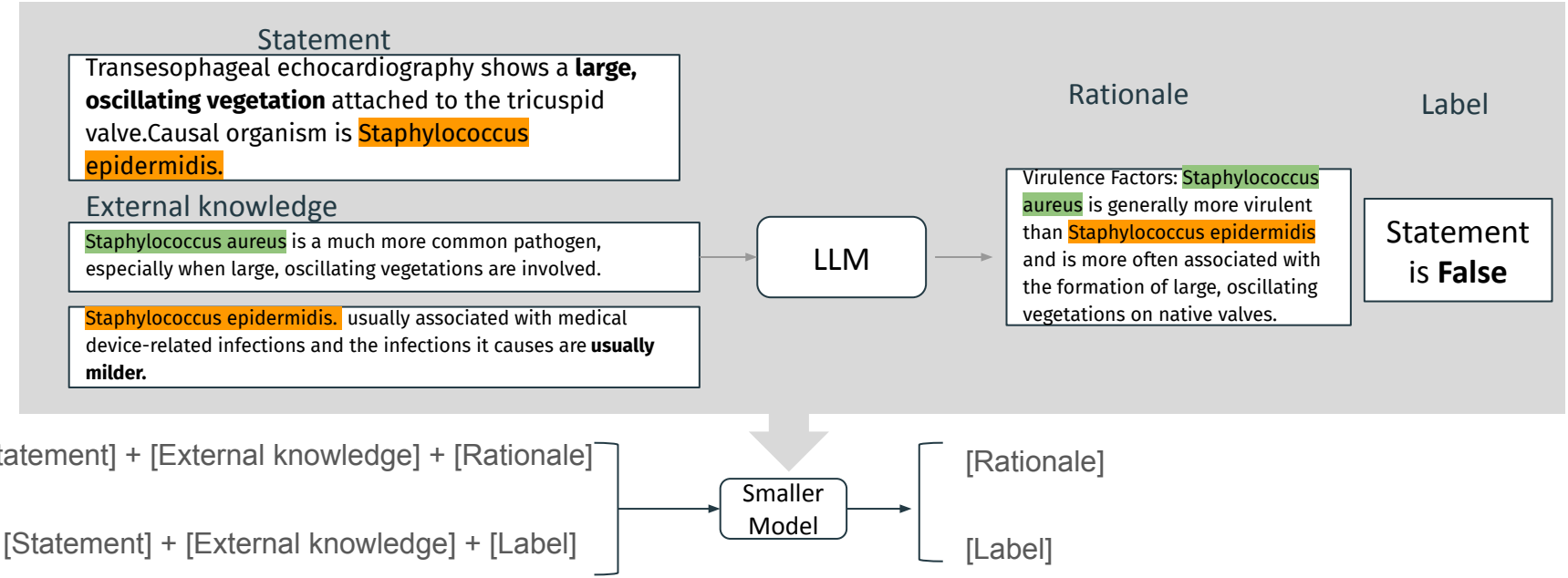


Zou, A., Zhang, Z., & Zhao, H. (2023). Decker: Double Check with Heterogeneous Knowledge for Commonsense Fact Verification. *arXiv preprint arXiv:2305.05921*.

Kang, M., Lee, S., Baek, J., Kawaguchi, K., & Hwang, S. J. (2024). Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, *36*.

Methods and external knowledge

1. External knowledge: **BioASQ** with medical knowledge question and answer.
2. Use LLM generate rationale to help binary classification.
3. Train a small model to distill the ability of absorb knowledge and binary classification.

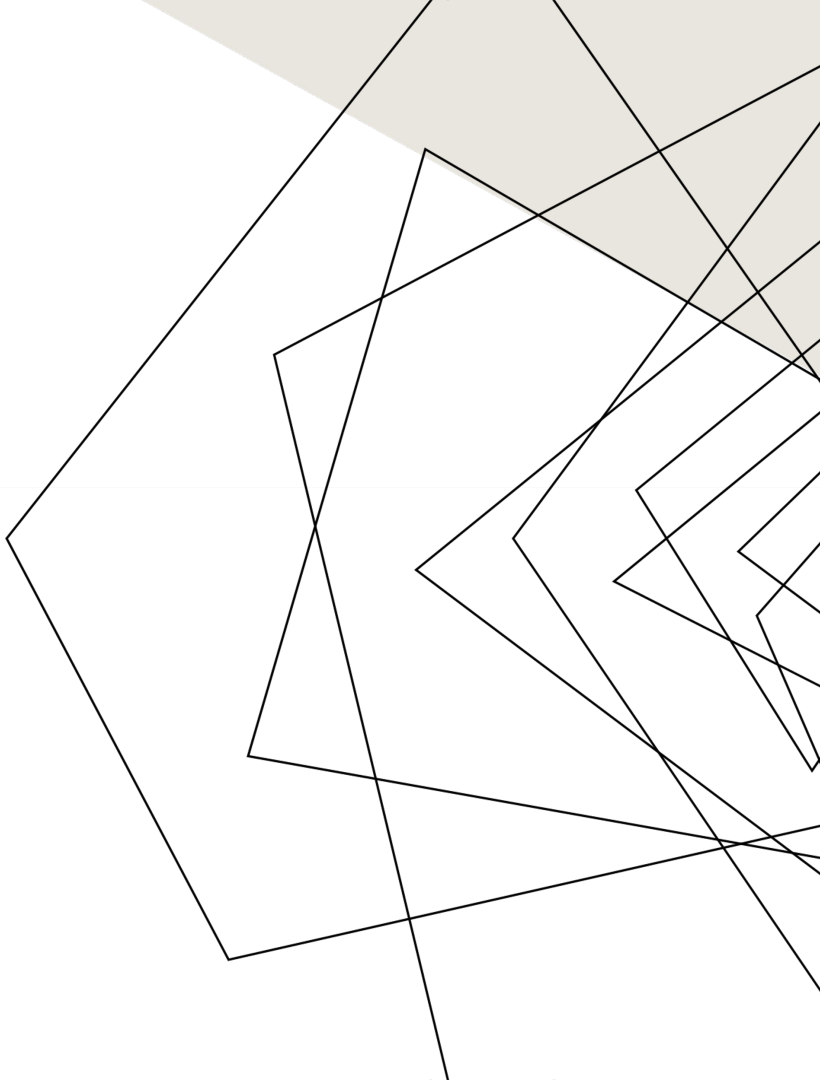


4. Yuning, Na, Ken, Yuelong

CONSTRUCTION OF CLASSIFIERS THAT CAN EXTRACT CHARACTER FEATURES

MOTIVATION

Our initial idea was to build a database of anime characters based on the language Cosplay, but since we didn't find anything that would work. So, we decided to make a classifier in the already constructed movie character database to allow us to take a line as input and be able to output the character most likely to say it.



RELATED WORKS

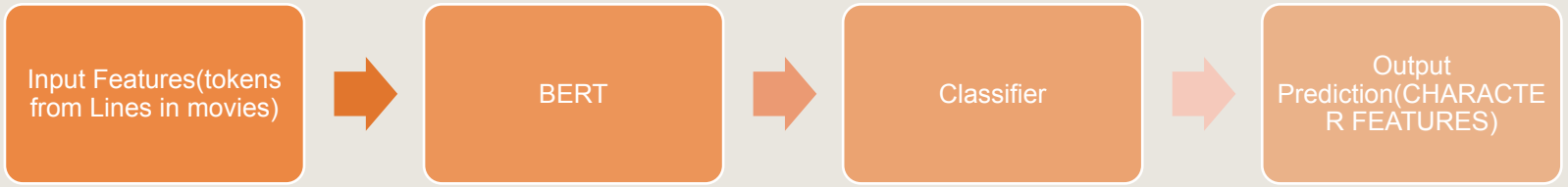
- **Neme, 2014:** Neme's work provided an approach to stylistic analysis and authorship attribution using self-organizing maps (SOMs), which used a dataset of over 140 texts in English and Spanish from 14 professional authors, 3 to 29 texts per author. Neme's team extract style features from text data and use the extracted features to train SOM.
- **Yang, 2017:**Yang's team referenced Neme's article. They also use feature selection to check the authorship attribution. The difference is that Yang's team did not use a dataset similar to Nemo's texts. They use a back propagation neural network optimized with particle swarm optimization (PSOBP) to attribute source code authorship accurately. In their conclusion, the classification accuracy (91.060%) is significantly improved.

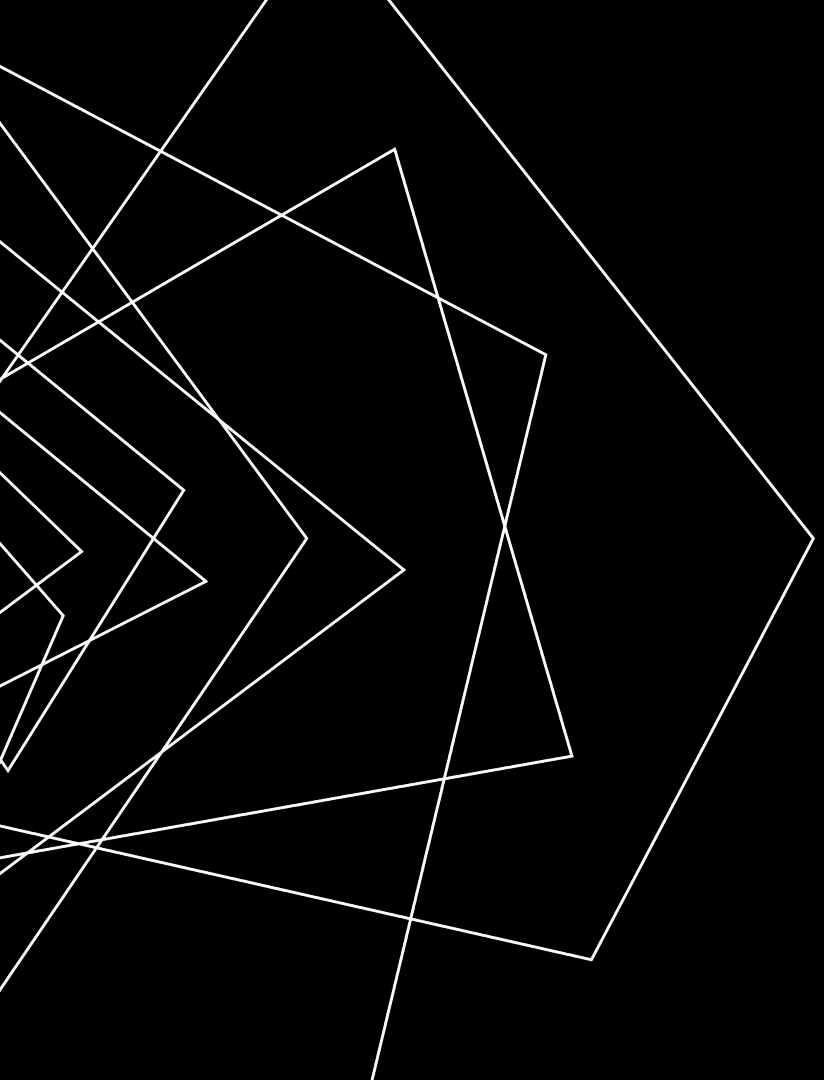
DATASET & EVALUATION

- We are planning to use Cornell Movie--Dialogs Corpus as our dataset. It has 220,579 conversational exchanges between 10,292 pairs of movie characters and involves 9,035 characters from 617 movies. This data set has a total of 304,713 utterances.
- We are planning to extract 3 to 4 series of movies to train our model. (e.g. Harry Poter, Transformers)
- Evaluation: accuracy, precision, recall, F1 score

APPROACHES

- BERT (Bidirectional Encoder Representations from Transformers)
- Why a BERT-based Model?
- Contextual Understanding: BERT and its variants are designed to understand the context of words in a sentence by considering the words that come before and after, which is crucial for distinguishing between characters based on their dialogue.
- Customization: Fine-tuning allows you to customize the pre-trained model to understand nuances and idiosyncrasies specific to movie characters' ways of speaking, which standard models might not capture.





THANK YOU

5. Sai, Deyasini, Shiva, Aparna

DETECTING TOXICITY IN HINGLISH MEMES

Agenda

AIM

Need for this Application

Approach

Outcomes

AIM!!

- Our project entails two significant contributions: Creating a dataset and the development of an innovative classification model, focusing on Hinglish memes.

- Our data set will have a curated collection of mixed coded memes of over 10 different classes.

- We aim to fit a binary classification model that can correctly classify Hinglish memes as either non-toxic or toxic.

NEED FOR THIS APPLICATION?

- Memes have become an integral part of online culture, but some carry harmful or toxic content.

- There are several models that can identify the toxic content in English memes but there are very few models that can do the same in mixed coded languages.

- So, we decided to choose a topic that can play a vital role in identifying harmful content, promoting responsible sharing, and safeguarding users.

APPROACH

- **Dataset Compilation:** Curate a strong Hinglish memes dataset for training a classifier, by collecting a variety of memes from different sources.
- **Model fitting:** Incorporating ML algorithms and Methodologies to examine features like language, imagery, context, and sentiment within memes.
- **Evaluation:** Evaluate it against the test set with metrics like Accuracy, Precision, F1 score etc...
- **Interpretability and Explainability**: Investigate model predictions to understand which features contribute to toxicity.

Proposed Outcomes

- Creation of a fresh one of a kind multilabel dataset.

- We are paving the way in the creation of a binary classification model especially for Hinglish meme analysis in conjunction with the dataset.

THANK YOU

- APARNA SRINIVASAN
- DEYASINI MITRA
- SAI VIVEK CHAVA
- SHIVA PATIBANDLA

6. Purva, Fatemeh, Ayush, Shayan

JestQuest: Identifying Humor in Text

Problem Statement and Objective

- **Challenge:** Distinguishing jokes from serious statements in digital content.
- **Goal:** Develop a system to accurately identify humorous text.
- **Challenges in NLP:**
  - Difficulty identifying humor in text.
  - Challenges with hate speech and irony.
- **Project Objective:**
  - Develop an annotated dataset for humor analysis.
  - Enhance NLP classifiers for humor recognition.

Evaluation and Ethical Considerations

- **Evaluation Metrics:**
  - Accuracy, precision, recall, F1 score.
  - Cross-validation techniques.
- **Ethical Considerations:**
  - Address bias and privacy concerns.
  - Mitigate misinterpretation of sensitive content.
  - Ensure cultural sensitivity.

Methodology

- **Data Collection:**
  - Curate diverse dataset.
  - Include public humor datasets.
  - Original annotations.
- **Annotation Process:**
  - Develop standardized annotation protocols.
  - Ensure consistency and accuracy in humor identification.

Future Scope

- **Future Directions:**
  - Further data annotation.
  - Model development and evaluation.

7. Nick, Arushi, Trung

BioLaySumm 2024: Evaluating LLMs for Biomedical Lay Summarization

Project motivation

Motivation:

- Biomedical publications are of interest to wide range of audiences.
- Technical and specialist language makes it hard for non-expert to understand.

Our goal: to develop an abstractive summarization model to generate lay summaries for non-technical people.

Dataset

PLOS, eLife: contains technical abstract and lay summary.

Dataset	# Train	# Dev	# Test
PLOS	24,773	1,376	142
eLife	4,346	241	142

Lay summary characteristics:

- PLOS: written by article's author, from 5 peer-reviewed journals
- eLife: written by editor in consultation with author, from eLife journals

Past works: Zero & Few Shot Prompting on GPT-3.5 Turbo

Zero shot prompting

Provide a lay summary of the following research abstract:
Abstract: "Diverse interactions among species within bacterial colonies lead to intricate spatiotemporal dynamics..."
Lay summary:

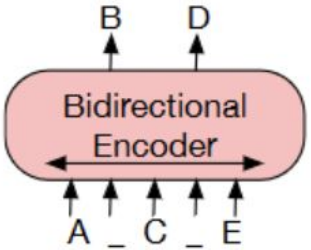
This research abstract discusses the impact of phenylketonuria (PKU) on Quality of Life (QoL). PKU is a genetic disorder that affects how the body processes a specific amino acid, phenylalanine. This study reviews existing research to understand how PKU influences QoL, considering factor ...

Few shot prompting

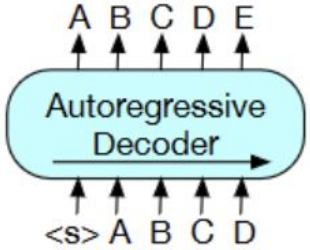
Provide a lay summary of the following research abstract:
Abstract: "Diverse interactions among species within bacterial colonies lead to intricate spatiotemporal dynamics..."
Lay summary: "Communities of bacteria and other microbes live in every ecosystem on Earth, including in soil..."
Abstract: "The ability to recognize foreign double-stranded (ds) DNA of pathogenic origin in the .."
Lay summary:

This research abstract discusses the impact of phenylketonuria (PKU) on Quality of Life (QoL). PKU is a genetic disorder that affects how the body processes a specific amino acid, phenylalanine. This study reviews existing research to understand how PKU influences QoL, considering factor ...

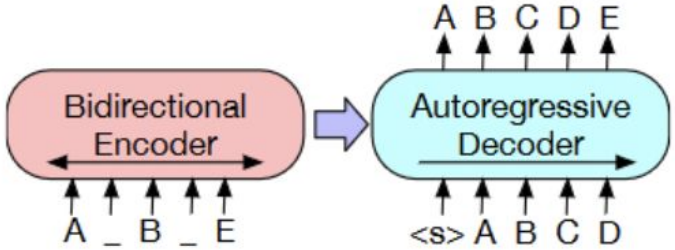
BART encoder-decoder model



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Approach

Utilizing GPT-4, Google Gemma-2B, Microsoft Orca-2-13B (Mitra et al., 2023), and BioGPT we will experiment with:

1. The construction of prompts that contain both context and instruction
2. Fine-tuning using the instruction tuning technique LoRA
3. Incorporating additional data using the LaySumm dataset and generating synthetic data

Evaluation

Evaluation is done for three criteria: relevance, readability, and factuality.

Relevance:
ROUGE1, ROUGE2, ROUGE-L, BERTScore

Readability:
Flesch-Kincaid Grade Level (FKGL), Dale-Chall Readability Score (DCRS), Coleman-Liau Index (CLI), Linsear Write Readability Formula (LENS)

Factuality:
AlignScore, SummaC

Thank you

Questions?

8. Kartik, Kasvitha, Brian, Owen

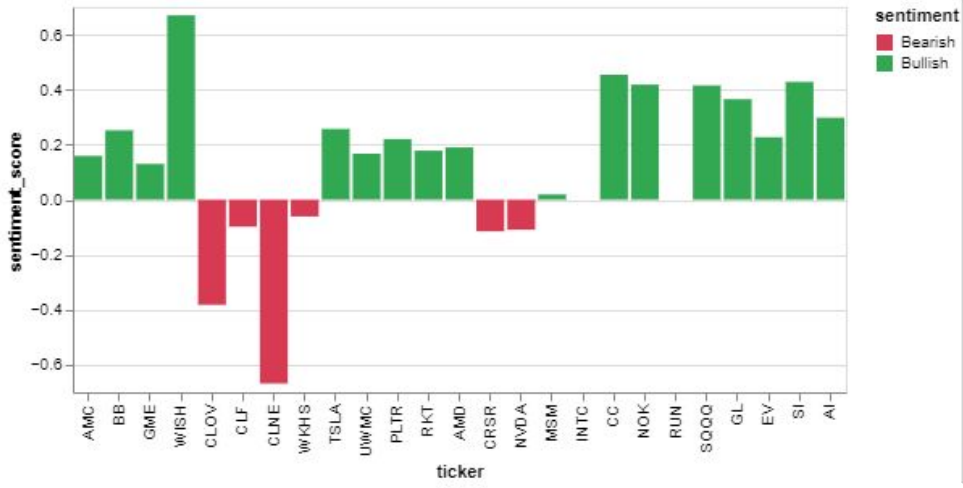
Context & Motivation

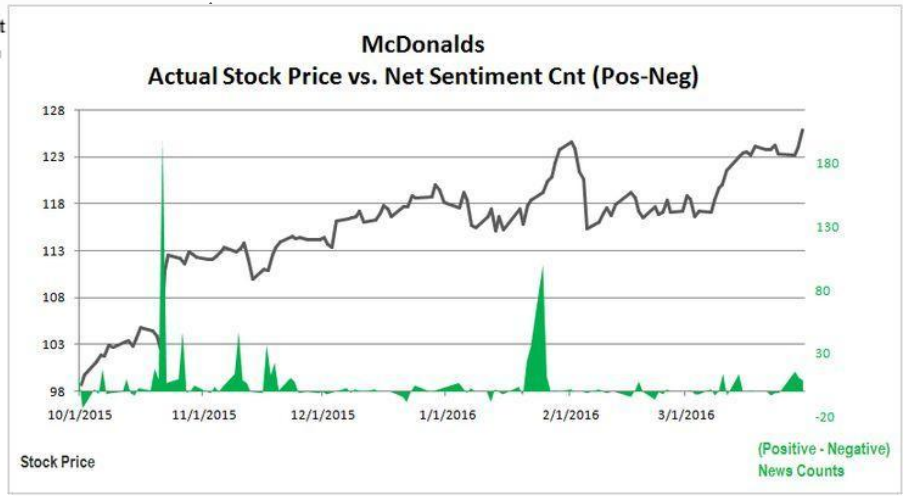
- Stock prices are volatile and hard to predict
- Influenced by many external factors
  - Public opinion
  - Company restructuring
  - General market trends
  - Etc.

- Accurately predicting stock movement can yield high returns
- Use financial news articles

Previous Work

1. BERT able to make sentiment analysis predictions using tweet data that directly correlated with NYSE market changes (Sebastian et al.)
2. BERT has also been trained to make accurate sentiment predictions





What makes us different?

- Does fake news or news from less-credible sources impact stock prices differently than verified news sources?

- Given that we know an article’s credibility level, can we better predict future stock prices?

Approach & Evaluation

2 Parts:

1. Can we predict the credibility of stock related news and tweets?
    - Train 3 models to predict the credibility of financial news
    - BERT model, RF model, and a kNN model
2. Can we combine the predictions of the above model with a more traditional sentiment analysis model to better predict stock price changes?
    - Train models that combine each of our credibility predictors with a BERT model for sentiment analysis to predict stock price change within a given time window
    - Compare models with a BERT model for sentiment analysis that does not use a credibility assessor

Datasets

- Tweet credibility labeled credible/not credible
- News article credibility labeled fake/real
- Historical Stock Price Data from Yahoo Finance API
- Stock related tweets from kaggle
- Stock related news dataset

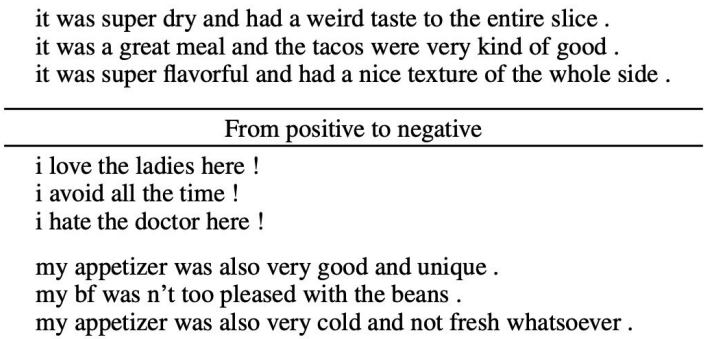
9. Noah, Annanya, Jayden, Xiaoyan

Machine translation with Style transfer

- Machine translation - well developed subfield of NLP
- Style transfer is less developed
  - Style: sentiment, gender of author, political slant
  - Very little work on doing both simultaneously
- Can we apply modern style transfer techniques to machine-translated text?

Related work

- Prabhumoye et al. 2018.
  - Friends with Michael
  - Style transfer in a single language using back translation
    - English -> French -> English - make a style-free latent representation of a sentence
    - Apply stylizing decoders to latent representation to give sentence with original semantic meaning and desired style
- Shen et al. 2017
  - Style transfer from non-parallel text
    - We have no parallel text
  - Develop cross aligned autoencoder



Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation .CoRR, abs/1804.09000.
Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 6833–6844, Red Hook, NY, USA. Curran Associates Inc.

Our approach and evaluation

- Employ a strategy like Prabhumoye et al. across languages, using the cross aligned autoencoder architecture as a generator
- chABSA dataset - Japanese sentiment classification dataset
- See how latent initial translations from English are
  - Hypothesis from Prabhumote et al.: translations to dissimilar languages will lose both more semantic meaning and style
  - Fine tune translator in an adversarial way to wash out style
- Evaluation metrics
  - Classification accuracy on style-transferred outputs
  - Human readability