Template: https://www.overleaf.com/project/67b7dcdfb9ae12009060dbdc

Proposal

Due Monday by 11:59pm
Points 5

A main component of this course is a final project (involving 1 to 3 students) on a topic in natural language processing focusing on the use of NLP in support of an empirical research question.

The proposal (due Mon Feb 24 by 11:59pm) should:

- outline the work you're going to undertake
- motivate its rationale as an interesting question worth asking
- assess its potential to contribute new knowledge by situating it within related literature in the scientific community. (cite 5 relevant sources)
- specify who is on the team, what each of your responsibilities are, and internal deadlines when those tasks will be completed. You should provide enough detail and specificity here that you can hold your teammates accountable for carrying out their share.

To turn in this assignment, please upload one pdf file per group. All reports should use the ACL style files, which are available as an Overleaf template.

Your proposal should be **1000 words [+/- 200 words]** long, **excluding the list of references at the end.** As you engage with other research papers, be sure to avoid inadvertently plagiarizing in anything you write -- review this helpful infographic on it by Emily Myers for a refresher. You, not an AI assistant, must be the author of any text you submit.

Midterm report

Due Mar 31 by 11:59pm
Points 10

The midterm report (due Mon, March 31 at 11:59pm) should detail your progress so far on your final project. Every project should have individual feedback at this point with expectations about what the midterm report should include. In general, at this stage, this includes:

-- Your literature review of related work should be complete; cite at least 10 sources that are immediately relevant to your work and contextualize how your proposed methods fit into the landscape of what's already been done. As a rough rule of thumb, you should probably be reading/skimming through 25-30 papers to find those 10 sources that matter. (A main part of the evaluation at this stage will be the thoroughness of this review.)

-- All of the data should ideally be collected.

-- A few preliminary experiments to test whether your proposed methods are likely to be feasible or if they need adjustment.

-- Detail an evaluation strategy for assessing the performance on your method.

-- As with the proposal, list concrete tasks you will carry out, who on your team will carry them out, and what the deadline is for them to so. Provide enough specificity that you can hold them accountable.

To turn in this assignment, please upload one pdf file per group. All reports should use the ACL style files, which are available as an Overleaf template

. The midterm report should be **2000 words [+/- 200 words]**, **excluding the list of references at the end.** As you engage with other research papers, be sure to avoid inadvertently plagiarizing in anything you write -- review this helpful infographic on it by Emily Myers for a refresher. You, not an AI assistant, must be the author of any text you submit.

Presentation

Due Apr 28 by 11:59pm
Points 5

Final project presentations will take place **in class** on Tuesday, April 29 and Thursday, May 1. For this session, prepare a single slide describing your project and be prepared to take questions from the audience (including the instructors). Your work should be relatively complete at this point. (You'll have time to incorporate any feedback you receive before your final project reports are due, but your experiments and analysis should be complete by this stage.)

General guidelines:

- You should be about 85% done with your work at this point (enough that the end is in sight but not necessarily finished).
- Your slide should outline basic information about the project goals. methods or findings so far.
- Aside from the presentation itself, a deliverable for this assignment will be a **single slide pdf**; submit that here on bCourses by 11:59pm on Monday April 28 (whether you present on Tuesday or Thursday).

Final report

Due May 12 by 11:59pm
Points 25

Final project report, due Monday, May 12 (11:59pm)

Your final project should be **3000 words [+/- 200 words long**, **excluding the list of references at the end**, single spaced in the ACL format. In addition to the paper, **you will also be submitting the code that you have written for this project so that we are able to reproduce the results**.

You'll be evaluated on the following criteria:

- Reproducibility. To what degree does the code submitted enable replication of results presented in the paper?

- Clarity. For the reasonably well-prepared reader, is it clear what was done and why? Is the paper well-written and well-structured?
- Originality. How original is the approach or problem presented in this paper? Does this paper break new ground in topic, methodology, or content? How exciting and innovative is the research it describes?
- Soundness. Is the technical approach sound and well-chosen? Second, can one trust the claims of the paper -- are they supported by proper experiments, proofs, or other argumentation?
- Substance. Does this paper have enough substance, or would it benefit from more ideas or results? Do the authors identify potential limitations of their work?
- Evaluation. To what extent has the application or tool been tested and evaluated?
- Meaningful comparison. Do the authors make clear where the presented system sits with respect to existing literature? Are the references adequate? Are the benefits of the system/application well-supported and are the limitations identified?
- Impact. How significant is the work described? Will novel aspects of the system result in other researchers adopting the approach in their own work?

By now, you'll have read several papers that are written in this format; for further examples, see:

https://aclanthology.org/events/acl-2024/

You'll note that many of the papers follow a standard format including an "Introduction," "Related Work," "Data," "Method," "Analysis," and "Conclusion".

To turn in this assignment, please upload:

- one pdf file per group containing your project report.
- **a zip file (.zip or .tar.gz) containing code and data** needed to reproduce your work. Do **not** submit any private information (including any API keys for LLM services you may have used -- be sure those are not hard-coded into your scripts).

All reports should use the ACL style files, which are available as an Overleaf template

. As you engage with other research papers, be sure to avoid inadvertently plagiarizing in anything you write -- review this helpful infographic on it by Emily Myers for a refresher. You, not an AI assistant, must be the author of any text you submit.

Topic	**Data Source**	**Function**	**Ref**	
Fake Information Detection	-News -Fake New(as pattern or xx)	Given information (could be social media post), check if its misinformation or not	Focus	Prem, Yunka Mimpi
JD keyword finder	Job Description	Given text of job description, return the keywords	Vendors: simplify, job scan	Prem Yunkai
Logical Reasoning - Cook book Recipe generation				Mimp Prem
Financial News Summarization/Sentiment	Social Media, News	- Given a stock name, should we invest or not - Stock Price will go up / down	https://www.researchgate.net/profile/Rodrigue-Andrawos/publication/361164679_NLP_in_Stock_Market_Prediction_A_Review/links/62a0885255273755ebdc1bd0/NLP-in-Stock-Market-Prediction-A-Review.pdf	Prem Mimp
Sustainability Report		1. check what the company with good rating wrote in their report 2. and if they actually did it		Mimp

Project Overview

Title: The Role and Effectiveness of Community Notes in Information Verification

Abstract: Community Notes is a crowdsourced fact-checking system designed to provide additional context and verification to online discussions. Currently in use on Twitter, it plays a significant role in combating misinformation by allowing users or contributors to add clarifying notes to tweets, while enabling the community to vote on whether the notes are helpful or not. However, building an open, participatory system comes with challenges, including resistance to manipulation, ensuring diverse perspectives, and maintaining high-quality contributions. This paper examines three key aspects of Community Notes: **analyzing Community Note data**, **the effectiveness and accuracy of the information provided**, and **its potential role in enhancing LLMs factuality**. By analyzing data on misinformation trends, cross-referencing note accuracy, and exploring AI integration, this research highlights the strengths and limitations of Community Notes. As major platforms like Facebook plan to implement similar systems in 2025, understanding the impact and scalability of Community Notes becomes increasingly crucial. This paper aims to provide insights into optimizing community-driven fact-checking for public discourse and AI applications. *<need to add what data we will use in this project as well as the scope of this project>*

- **General Topic Reference**
    - Can Community Notes Replace Professional Fact-Checkers?
      19 Feb 2025 https://arxiv.org/html/2502.14132v1
    - Community notes increase trust in fact-checking on social media 2024 May 31 https://pmc.ncbi.nlm.nih.gov/articles/PMC11212665/
    - Jul 2023 https://publications.iadb.org/en/reducing-misinformation-role-confirmation-frames-fact-checking-interventions
- **Aspect 1 analyzing Community Note data**
    - Community Notes vs. Snoping:How the Crowd Selects Fact-Checking Targets on Social Media *https://ojs.aaai.org/index.php/ICWSM/article/view/31387/33547*
- **Aspect 2 the effectiveness and accuracy of the information provided**
    - *Scope to politifact, political related*
    - Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes https://arxiv.org/html/2406.12444v1
    - Report says crowd-sourced fact checks on X fail to address flood of US election misinformation https://apnews.com/article/x-musk-twitter-misinformation-ccdh-0fa4fec0f703369b93be248461e8005d
    -
    -
- *Aspect 3* **potential role in enhancing LLMs factuality**

- Unpacking DPO and PPO: Disentangling Best Practices for Learning from Preference Feedback https://arxiv.org/abs/2406.09279
    - It mentions about how factuality accuracy score isn't improve a lot when It comes to preference feedback evaluation
- Evaluating the Factuality of Large Language Models using Large-Scale Knowledge Graphs https://arxiv.org/pdf/2404.00942
    - the factual accuracy score improved for certain models, particularly within the LLaMA series, but with some trade-offs
- https://medium.com/@techsachin/openfactcheck-a-unified-framework-for-factuality-evaluation-of-llms-d88f2946ca94
-

Background and Motivation

Fact-checking is essential for verifying the accuracy of information, particularly on issues like politics, public health, and climate change. This need is especially acute on social media platforms such as Twitter/X and Facebook, where misinformation can spread rapidly due to several factors such as anonymous posting, people are more likely to retweet false or novel information [1], algorithms that prioritize engagement over accuracy, and lack of verification mechanisms. Traditional fact-checking relies on human experts who cross-reference claims with credible sources to assess their validity. Organizations such as PolitiFact, Snopes, and FactCheck.org have played an essential role in mitigating misinformation. However, human fact-checking can be time-consuming, limited in scale, and can be influenced by personal biases. To address these limitations, many tools like GPTzero or Originality.AI using AI-driven fact-checking have emerged, leveraging machine learning and natural language processing (NLP) to analyze and verify information more efficiently. However, AI-based tools also have their challenges, including biases in training data, susceptibility to adversarial misinformation, difficulty understanding nuanced contexts, and the risk of reinforcing false narratives if trained on unreliable sources.

The Rise of Community Notes:

Community Notes, originally launched as Birdwatch by Twitter in January 2021, is a community-based fact-checking system that empowers users to add context to posts they believe contain false or misleading information. Rather than being published immediately, these notes undergo a review process where contributors vote on their usefulness. Only when the votes come from a diverse range of perspectives does the note become public, offering additional context that helps users assess the original content. Research has demonstrated that this approach can achieve high accuracy—studies on COVID-related notes have shown accuracy rates as high as 97%—and it can significantly reduce the spread of misinformation [2] However, research also indicates that the review process can be slow, sometimes delaying the publication of notes until after misinformation has already circulated. In 2024, with Zuckerberg noting, “We’ve reached a point where it’s just too many mistakes and too much censorship,” platforms like Facebook and Instagram will transition from third-party fact-checking to using Community Notes by 2025. This shift is important because it marks a move toward greater transparency and community involvement in moderating online content, which could lead to a more accountable and responsive information environment.

Therefore, as Community Notes play a critical role in combating misinformation and promoting transparency, our work focuses on harnessing their potential to enhance fact-checking processes. We examine three key aspects: analyzing Community Note data to identify misinformation trends, evaluating the effectiveness and accuracy of the provided context, and exploring how these insights can be integrated with AI to improve the factuality of large language models. With major platforms like Facebook set to adopt similar systems, It's vital to understand how Community Notes perform and expand can help us enhance community-based fact-checking, ensuring that our public discourse remains informed and that AI tools work with accurate data.

[1] Study: On Twitter, false news travels faster than true stories, https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308

[2] Do Community Notes work?

https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/14/do-community-notes-work/

[3] Meta is getting rid of fact checkers. Zuckerberg acknowledged more harmful content will appear on the platforms now

https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html

Related Work

Methodology

Approach:

1) Collecting and processing Data
    a) Community notes
    b) PolitiFact (2024 data) focus on Election
2) Learning About community note by analyzing data
    a) Topic Analysis: using topic modeling like [xx]
    b) xxx
3) Community Notes Accuracy
    a) If we do politifact
        i) Use 50% false information, 50%
    b) LLMs
4)

Data Sources:

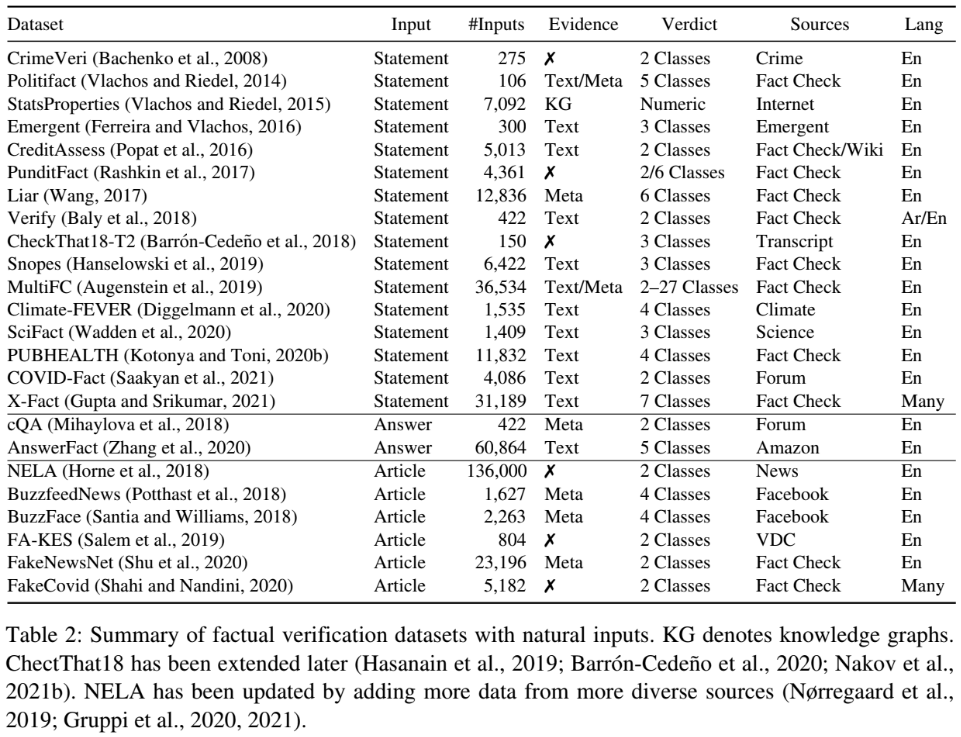
- Twitter community note
- Existing DB like FakeNewsNet https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/UEMMHS, https://github.com/KaiDMML/FakeNewsNet?tab=readme-ov-file , Politifact

Techniques & Tools:

- Aspect1 Analysis

Evaluation Metrics:

- Aspect1 - No need?
- Aspect2 - Accuracy to Already Checked Data



Dataset	Input	#Inputs	Evidence	Verdict	Sources	Lang
CrimeVeri (Bachenko et al., 2008)	Statement	275	✗	2 Classes	Crime	En
Politifact (Vlachos and Riedel, 2014)	Statement	106	Text/Meta	5 Classes	Fact Check	En
StatsProperties (Vlachos and Riedel, 2015)	Statement	7,092	KG	Numeric	Internet	En
Emergent (Ferreira and Vlachos, 2016)	Statement	300	Text	3 Classes	Emergent	En
CreditAssess (Popat et al., 2016)	Statement	5,013	Text	2 Classes	Fact Check/Wiki	En
PunditFact (Rashkin et al., 2017)	Statement	4,361	✗	2/6 Classes	Fact Check	En
Liar (Wang, 2017)	Statement	12,836	Meta	6 Classes	Fact Check	En
Verify (Baly et al., 2018)	Statement	422	Text	2 Classes	Fact Check	Ar/En
CheckThat18-T2 (Barrón-Cedeño et al., 2018)	Statement	150	✗	3 Classes	Transcript	En
Snopes (Hanselowski et al., 2019)	Statement	6,422	Text	3 Classes	Fact Check	En
MultiFC (Augenstein et al., 2019)	Statement	36,534	Text/Meta	2–27 Classes	Fact Check	En
Climate-FEVER (Diggelmann et al., 2020)	Statement	1,535	Text	4 Classes	Climate	En
SciFact (Wadden et al., 2020)	Statement	1,409	Text	3 Classes	Science	En
PUBHEALTH (Kotonya and Toni, 2020b)	Statement	11,832	Text	4 Classes	Fact Check	En
COVID-Fact (Saakyan et al., 2021)	Statement	4,086	Text	2 Classes	Forum	En
X-Fact (Gupta and Srikumar, 2021)	Statement	31,189	Text	7 Classes	Fact Check	Many
cQA (Mihaylova et al., 2018)	Answer	422	Meta	2 Classes	Forum	En
AnswerFact (Zhang et al., 2020)	Answer	60,864	Text	5 Classes	Amazon	En
NELA (Horne et al., 2018)	Article	136,000	✗	2 Classes	News	En
BuzzfeedNews (Potthast et al., 2018)	Article	1,627	Meta	4 Classes	Facebook	En
BuzzFace (Santia and Williams, 2018)	Article	2,263	Meta	4 Classes	Facebook	En
FA-KES (Salem et al., 2019)	Article	804	✗	2 Classes	VDC	En
FakeNewsNet (Shu et al., 2020)	Article	23,196	Meta	2 Classes	Fact Check	En
FakeCovid (Shahi and Nandini, 2020)	Article	5,182	✗	2 Classes	Fact Check	Many

Table 2: Summary of factual verification datasets with natural inputs. KG denotes knowledge graphs. ChectThat18 has been extended later (Hasanain et al., 2019; Barrón-Cedeño et al., 2020; Nakov et al., 2021b). NELA has been updated by adding more data from more diverse sources (Nørregaard et al., 2019; Gruppi et al., 2020, 2021).

- Aspect3 - Accuracy with Benchmark

Team & Responsibilities

Bhirajaya Bhumimars

Supakarn Bunlert

Yunkai Li

Timeline

References

Role of Community Notes

Meta/ Article

Meta is ditching fact checkers for X-style community notes. Will they work?
25 Jan 2025
https://www.bbc.com/news/articles/c4g93nvrdz7o

Fact Checking vs Community Notes: The Unintentional case for Media Literacy
January 23, 2025
https://www.linkedin.com/pulse/fact-checking-vs-community-notes-unintentional-case-media-dejong-rtdgf/

Community Notes and its Narrow Understanding of Disinformation
Feb 3, 2025
https://www.techpolicy.press/community-notes-and-its-narrow-understanding-of-disinformation/

Fact-checkers are among the top sources for X's Community Notes, study reveals
February 20, 2025
https://www.poynter.org/ifcn/2025/fact-checkers-contribute-improve-community-notes-x/

Papers

Can Community Notes Replace Professional Fact-Checkers?
19 Feb 2025
https://arxiv.org/html/2502.14132v1

Community notes increase trust in fact-checking on social media
2024 May 31
https://pmc.ncbi.nlm.nih.gov/articles/PMC11212665/

Who Checks the Checkers? Exploring Source Credibility in Twitter's Community Notes
https://arxiv.org/html/2406.12444v1