Headline Grouping: A Challenging NLU Task

Philippe Laban, Lucas Bandarkar, Marti A. Hearst
UC Berkeley

NAACL 2021 - Video Talk

New Challenges for NLU

Recent progress on NLU has surpassed human performance:

- **Paraphrase Identification** (MRPC): + 8 F-1 (compared to human performance)

- **Question Answering** (SQuAD): + 4 F-1

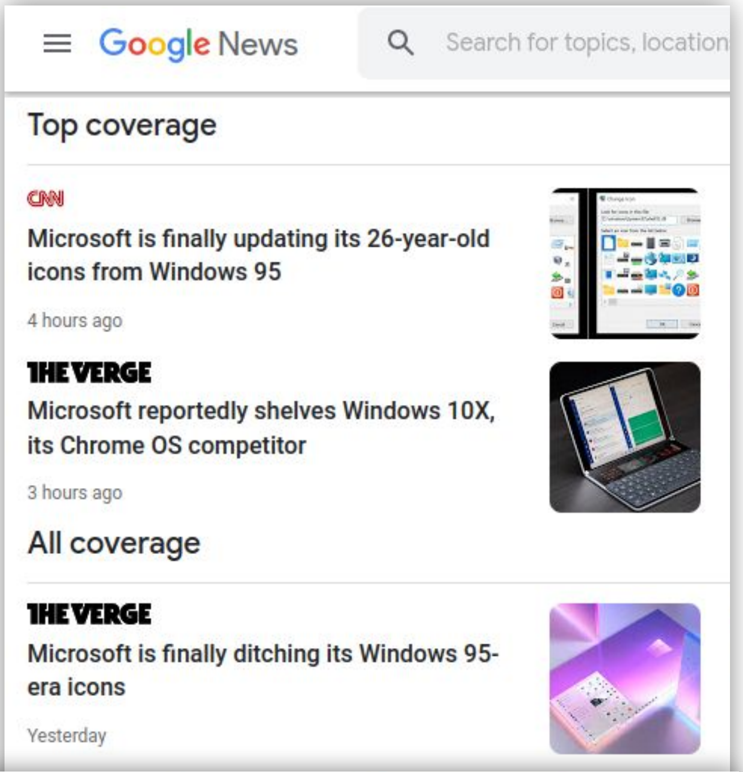
- **Textual Similarity** (STS-B): +1 F-1

Need more challenging NLU tasks

Headline Grouping for News Aggregation

News Aggregators group headlines to present *diverse coverage* for events.

Broader news coverage can help news readers form their more nuanced opinion.



Top coverage

CNN

Microsoft is finally updating its 26-year-old icons from Windows 95

4 hours ago

THE VERGE

Microsoft reportedly shelves Windows 10X, its Chrome OS competitor

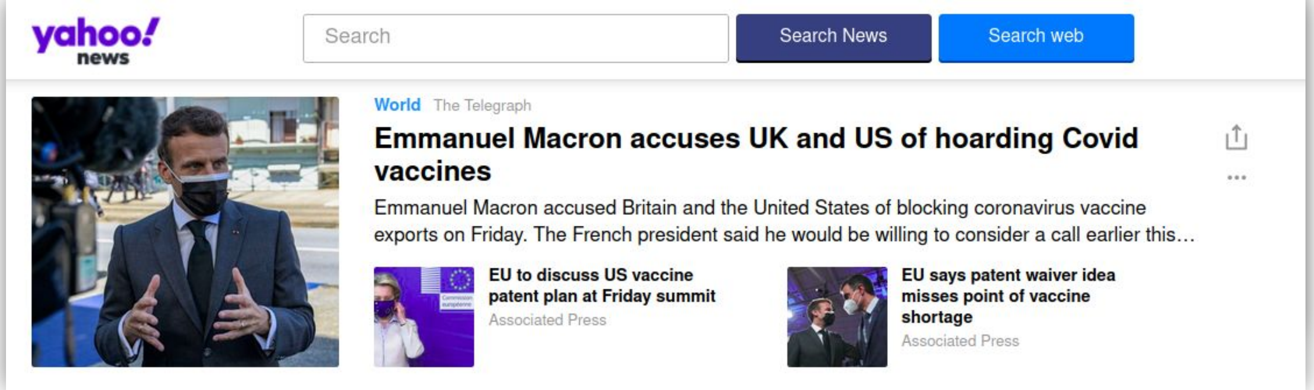
3 hours ago

All coverage

THE VERGE

Microsoft is finally ditching its Windows 95-era icons

Yesterday



World The Telegraph

Emmanuel Macron accuses UK and US of hoarding Covid vaccines

Emmanuel Macron accused Britain and the United States of blocking coronavirus vaccine exports on Friday. The French president said he would be willing to consider a call earlier this…

EU to discuss US vaccine patent plan at Friday summit

Associated Press

EU says patent waiver idea misses point of vaccine shortage

Associated Press

Headline Grouping Task

Challenges of Headline Grouping

Method	Percentage
Headlines differ in level of detail	37 %

NASA delays work on Moon rocket during virus pandemic

vs.

Nasa’s Moon plans take a hit

Challenges of Headline Grouping

Method	Percentage
Headlines differ in level of detail	37 %
Headlines are exact paraphrases	30 %

Equifax takes web page offline after reports of new cyber attack

vs.

Equifax takes down web page after reports of new hack

Challenges of Headline Grouping

Method	Percentage
Headlines differ in level of detail	37 %
Headlines are exact paraphrases	30 %
Headlines differ in aspect of focus	26 %

Astronauts to Get Thanksgiving Feast in Space

vs.

A Brief History of Thanksgiving Turkey in Space

Challenges of Headline Grouping

Method	Percentage
Headlines differ in level of detail	37 %
Headlines are exact paraphrases	30 %
Headlines differ in aspect of focus	26 %
Headlines contain humor, puns, etc.	7 %

New privacy law forces some U.S. media offline in Europe

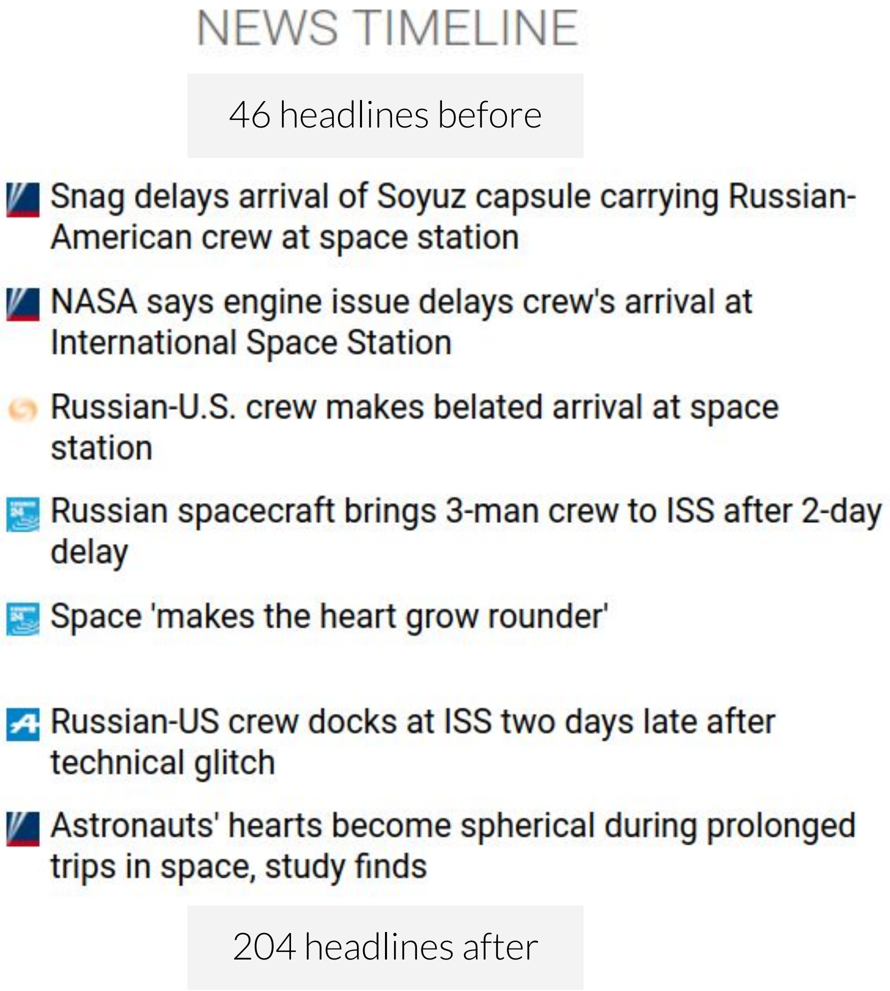
vs.

US websites blacked out in Europe on 'Happy GDPR Day'

Creating HLGD
(HeadLine Grouping Dataset)

HLGD consists of annotated **news timelines**.

Timeline: a chronological list of headlines covering a common story over time.



NEWS TIMELINE

46 headlines before

Snag delays arrival of Soyuz capsule carrying Russian-American crew at space station

NASA says engine issue delays crew's arrival at International Space Station

Russian-U.S. crew makes belated arrival at space station

Russian spacecraft brings 3-man crew to ISS after 2-day delay

Space 'makes the heart grow rounder'

Russian-US crew docks at ISS two days late after technical glitch

Astronauts' hearts become spherical during prolonged trips in space, study finds

204 headlines after

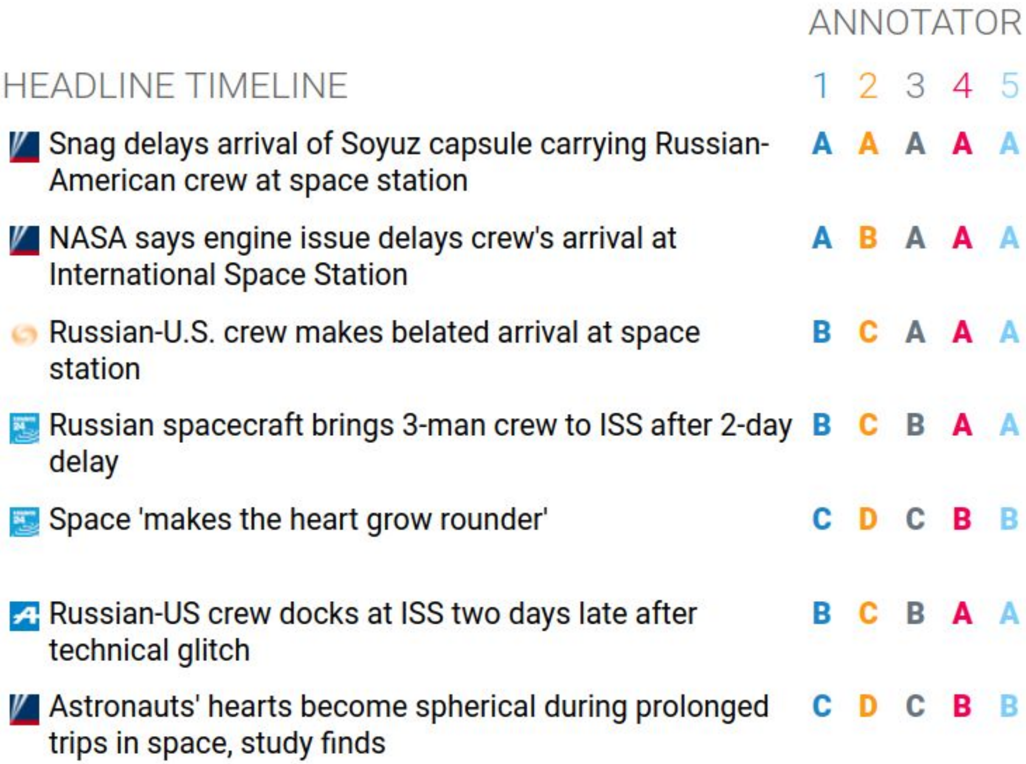
Creating HLGD (Headline Grouping Dataset)

10 news timelines with diverse topics and geography.

Timeline Name	# Headlines	
Tunisia Protests	111	TRAIN
Ireland Abortion Vote	180	
Ivory Coast Army Mutiny	128	
International Space Station	257	
US Bird Flu Outbreak	79	
Human Cloning	119	
Facebook Privacy Scandal	194	VALIDATION
Equifax Breach	159	
Brazil Dam Disaster	273	TEST
Wikileaks Trials	180	

Creating HLGD

Each timeline is annotated by **5 annotators**



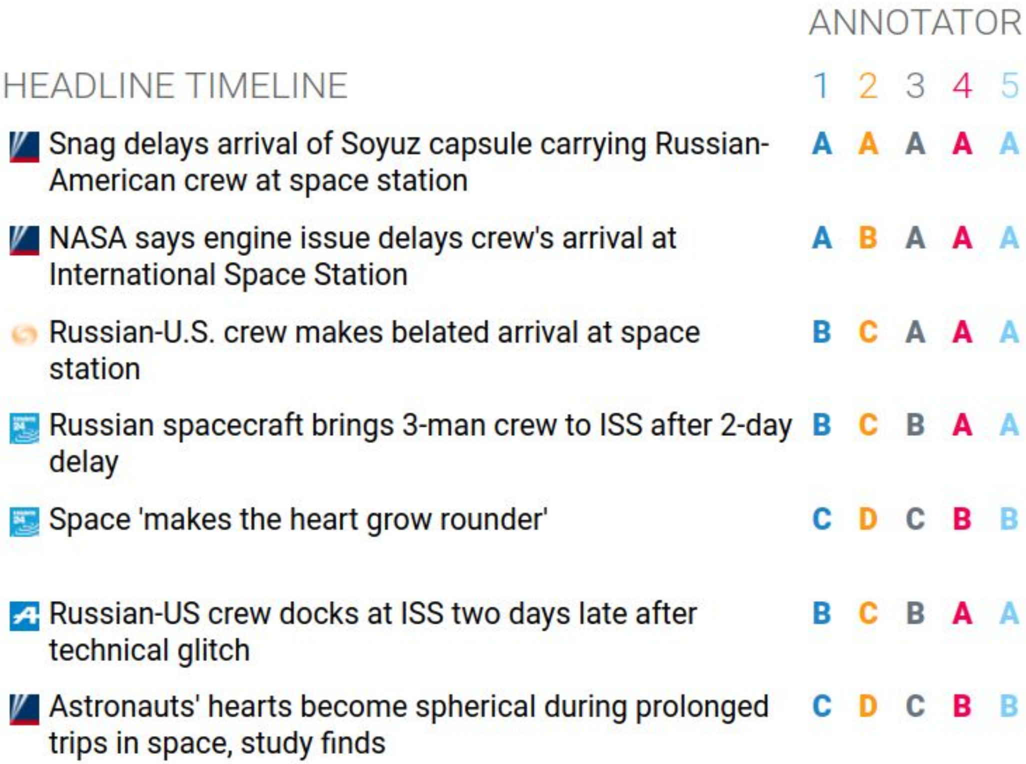
	ANNOTATOR				
HEADLINE TIMELINE	1	2	3	4	5
Snag delays arrival of Soyuz capsule carrying Russian-American crew at space station	A	A	A	A	A
NASA says engine issue delays crew's arrival at International Space Station	A	B	A	A	A
Russian-U.S. crew makes belated arrival at space station	B	C	A	A	A
Russian spacecraft brings 3-man crew to ISS after 2-day delay	B	C	B	A	A
Space 'makes the heart grow rounder'	C	D	C	B	B
Russian-US crew docks at ISS two days late after technical glitch	B	C	B	A	A
Astronauts' hearts become spherical during prolonged trips in space, study finds	C	D	C	B	B

Creating HLGD

Each timeline is annotated by **5 annotators**

Inter-annotator Agreement: 0.814

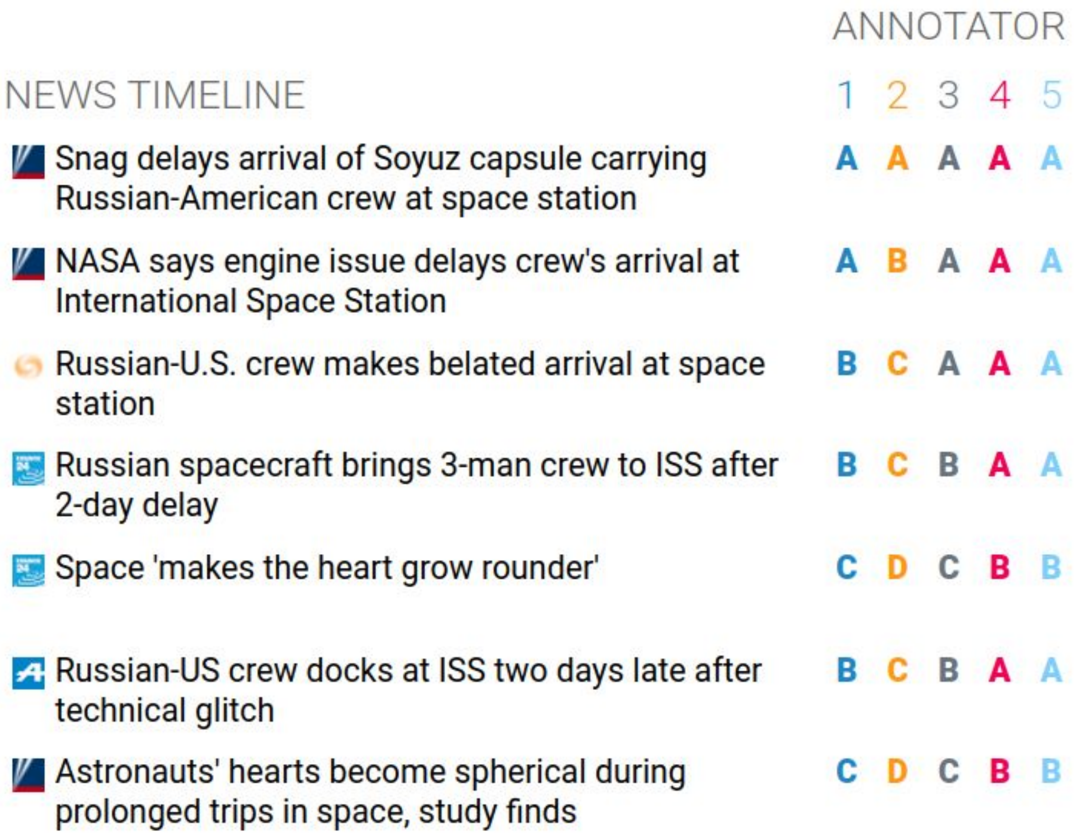
(using adjusted Mutual Information)



	ANNOTATOR				
HEADLINE TIMELINE	1	2	3	4	5
Snag delays arrival of Soyuz capsule carrying Russian-American crew at space station	A	A	A	A	A
NASA says engine issue delays crew's arrival at International Space Station	A	B	A	A	A
Russian-U.S. crew makes belated arrival at space station	B	C	A	A	A
Russian spacecraft brings 3-man crew to ISS after 2-day delay	B	C	B	A	A
Space 'makes the heart grow rounder'	C	D	C	B	B
Russian-US crew docks at ISS two days late after technical glitch	B	C	B	A	A
Astronauts' hearts become spherical during prolonged trips in space, study finds	C	D	C	B	B

Creating HLGD

Create a **global group** with majority vote and clustering



	ANNOTATOR					GLOBAL
NEWS TIMELINE	1	2	3	4	5	GROUP
Snag delays arrival of Soyuz capsule carrying Russian-American crew at space station	A	A	A	A	A	A
NASA says engine issue delays crew's arrival at International Space Station	A	B	A	A	A	A
Russian-U.S. crew makes belated arrival at space station	B	C	A	A	A	A
Russian spacecraft brings 3-man crew to ISS after 2-day delay	B	C	B	A	A	A
Space 'makes the heart grow rounder'	C	D	C	B	B	B
Russian-US crew docks at ISS two days late after technical glitch	B	C	B	A	A	A
Astronauts' hearts become spherical during prolonged trips in space, study finds	C	D	C	B	B	B

HLGD Classification

- Pairs of headlines in a timeline are either:
  - In the same global group **label = 1**
  - In different global groups **label = 0**

HLGD Classification

- Pairs of headlines in a timeline are either:
  - In the same global group **label = 1**
  - In different global groups **label = 0**

⚠ *Without further filtering:* ***large class imbalance***
(40 negatives for 1 positive)

HLGD Classification

- Pairs of headlines in a timeline are either:
  - In the same global group **label = 1**
  - In different global groups **label = 0**

Observation: 98% of positive headline pairs are published within **4 days** of each other.

HLGD Classification

- Pairs of headlines in a timeline are either:
  - In the same global group **label = 1**
  - In different global groups **label = 0**

Idea: Keep only negative pairs that are published within four days or less. Filtering out “easy” negatives.

HLGD Classification

- Pairs of headlines in a timeline are either:
  - In the same global group **label = 1**
  - In different global groups **label = 0**
  - Remove all negative pairs published more than four days apart

Final HLGD dataset

20k pairs (1-5 imbalance)

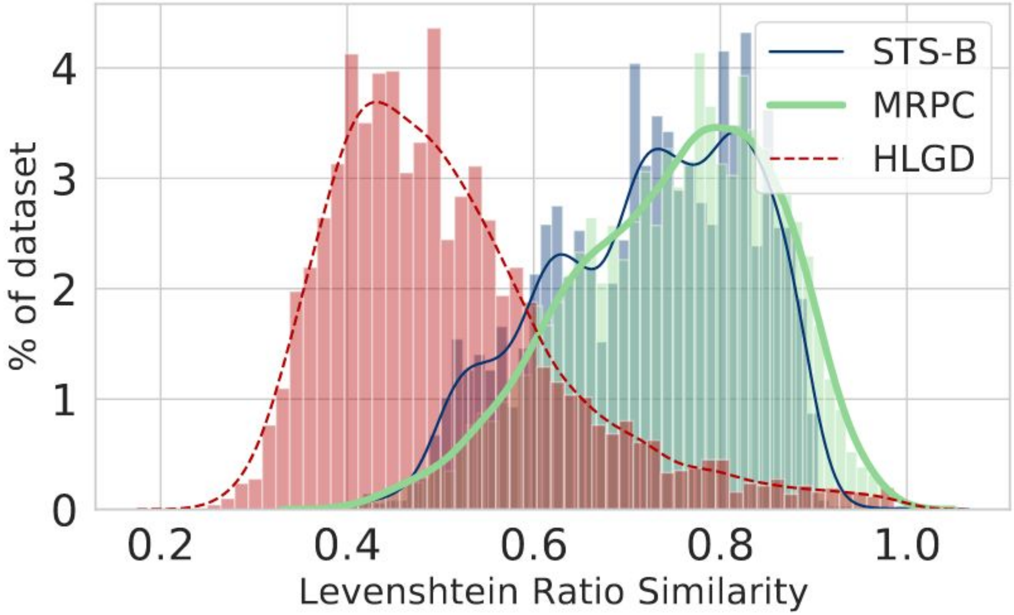
How does HLGD compare to other NLU datasets?

HLGD vs. Similar NLU Datasets

Headline Grouping is a binary classification task on an unordered sentence pair.

It is most similar to Paraphrase Identification and Textual Similarity tasks.

HLGD vs. Similar NLU Datasets



Distribution of **positive pairs** in each dataset

Challenge: Headlines can be in the same group while being syntactically distant

Challenge Settings

Which metadata can I use to make predictions on HLGD?

Challenge 1: Headline-only

Challenge 2: Headline + Date

Challenge 3: Headline + Date + Other

Baseline & Human Performance

- **Syntactic-Only:** 0.49 F-1
  - *Choose best threshold in Levenshtein ratio on validation set*

- **Time-only:** 0.59 F-1
  - *Choose best threshold in publication date difference on validation set*

- **Human-performance:** 0.90 F-1
  - *Obtained with independent 6th annotators on validation and test sets*

What if I finetune
a Transformer?

Directly Training on HLGD Pairs

- **Electra Finetune:** 0.80 F-1
  - *Model sees:* `Headline 1 <sep> Headline 2`

Using only headline pairs to make the task most similar to other Text Pair classification tasks (NLI, PI).

Directly Training on HLGD Pairs

- **Electra Finetune:** 0.80 F-1
    - *Model sees:* `Headline 1 <sep> Headline 2`
- **Electra Finetune + time:** 0.83 F-1
    - *Model sees:* `Headline 1 <sep> Headline 2 + publication day difference`

Adding publication date information helps increase performance by ~0.03 F-1.

Directly Training on HLGD Pairs

- **Electra Finetune:** 0.80 F-1
  - *Model sees:* `Headline 1 <sep> Headline 2`

- **Electra Finetune + time:** 0.83 F-1
  - *Model sees:* `Headline 1 <sep> Headline 2 + publication day difference`

- **Electra Content Finetune:** 0.73 F-1
  - *Model sees:* `Content 1 <sep> Content 2`

Surprisingly, using article's full content lowers instead of headlines lowers performance.

Can we use a Generator to Zero-Shot this task?

Could these headlines be swapped?

(while keep the body of the text constant)

ARTICLE 1

Tunisia Plans Social Reforms After Wave of Anti-Austerity Protests

Tunisia's government has announced a new package of social reforms worth nearly $70 million. The North African country has been rocked by protests ahead of the seventh anniversary of the Arab Spring uprising.

The Tunisian government on Saturday announced a social reforms package aimed at improving care for the needy and increasing access to health care following a wave of anti-austerity protests.

ARTICLE 2

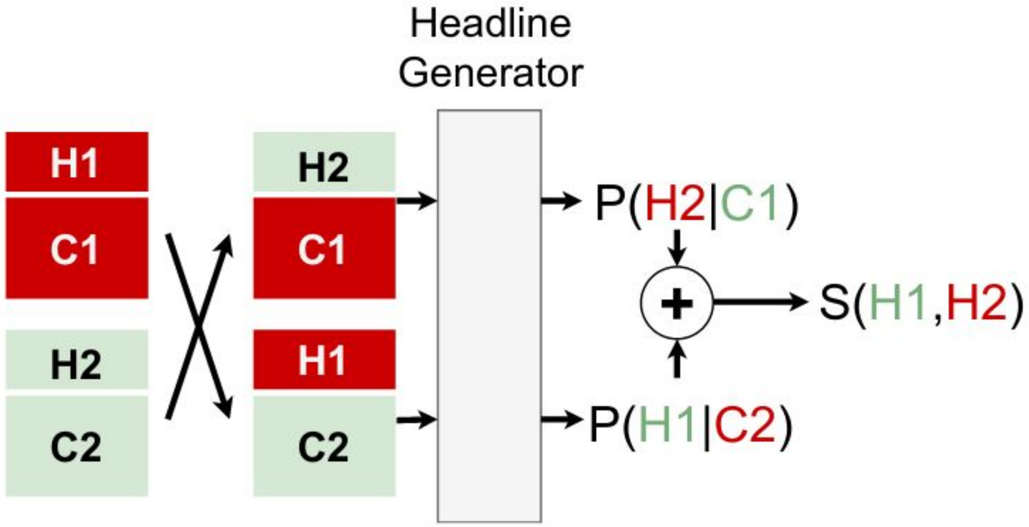
Tunisia protests: Government announce reforms after unrest

There were fresh protests on Sunday, the seventh anniversary of the ousting of President Zine al-Abidine Ben Ali.

Emergency government meetings have been held in response to the protests, which have seen more than 800 arrests.

President Beji Caid Essebsi visited a district of Tunis on Sunday, saying he understood the people's suffering.

Headline Generator Swap Model



Swap the headlines of two articles, score the swap with a headline generator to decide if headlines are in the same group.

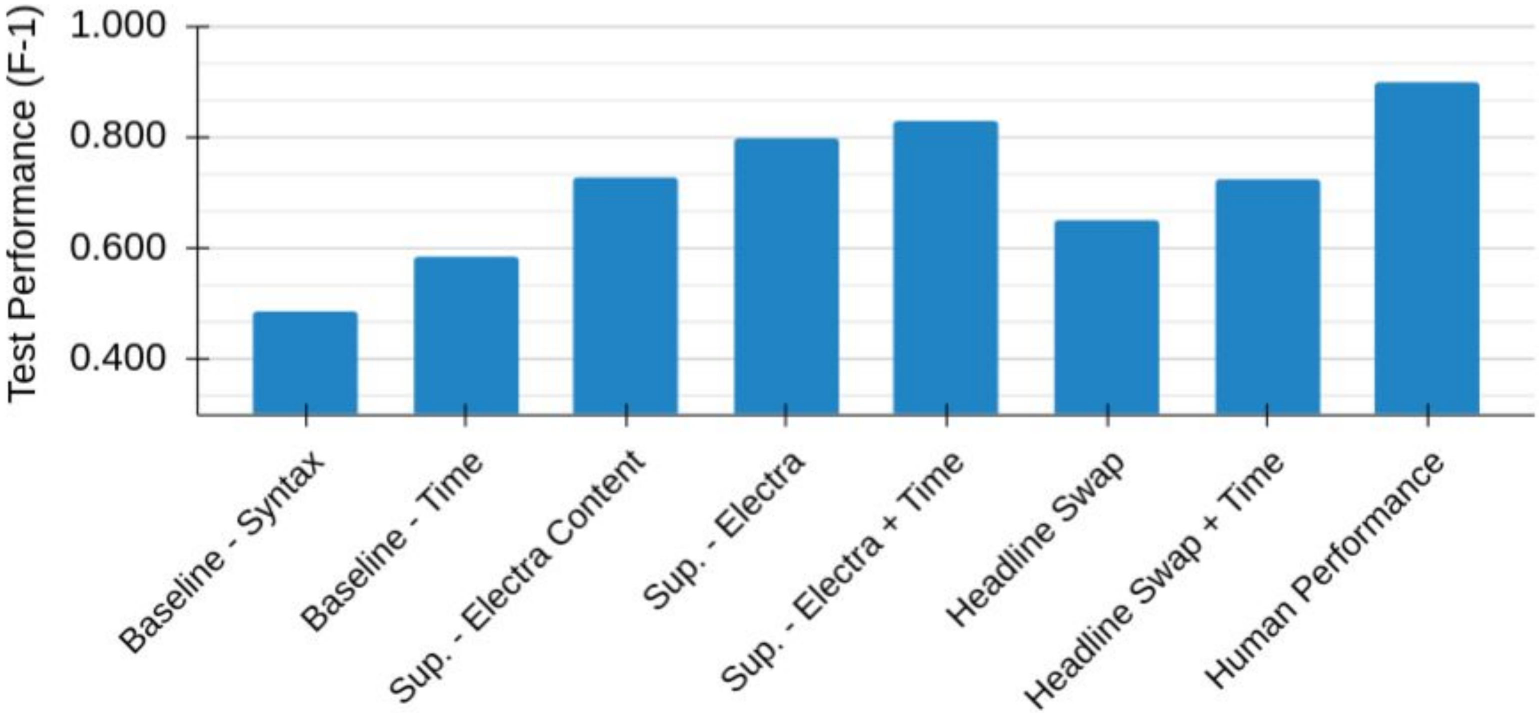
Headline Generator Swap Model

- **Headline Gen. Swap**: 0.651 F-1
  - *Model considers scores the swap, choose best threshold using validation set*

- **Headline Gen. Swap + time**: 0.722 F-1
  - *Multiplying score by publication day difference, choosing a different threshold on validation set*

With no training, performance is competitive with **supervised** models.

Compiled Results



F-1 Performance of the various models presented. (1) Supervised models achieve best automatic performance, (2) time information helps but isn't enough on its own.

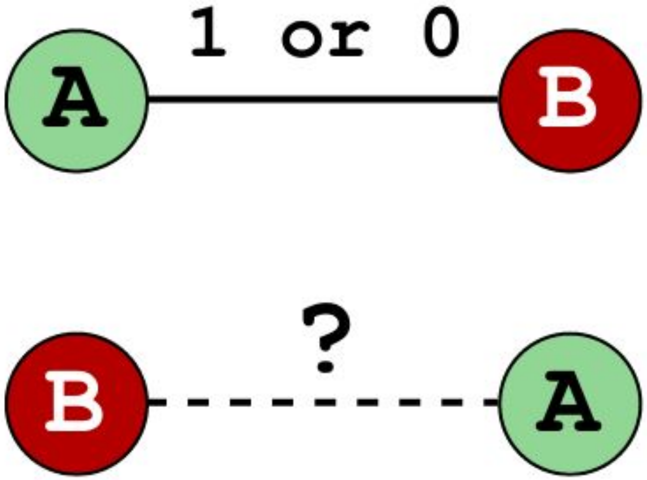
Model Consistency Analysis

The Headline Grouping task assumes some properties.
Are models consistent with these properties?

Model Commutativity

The model processes headlines in an **arbitrary order**.

Does this order have an impact on model prediction?

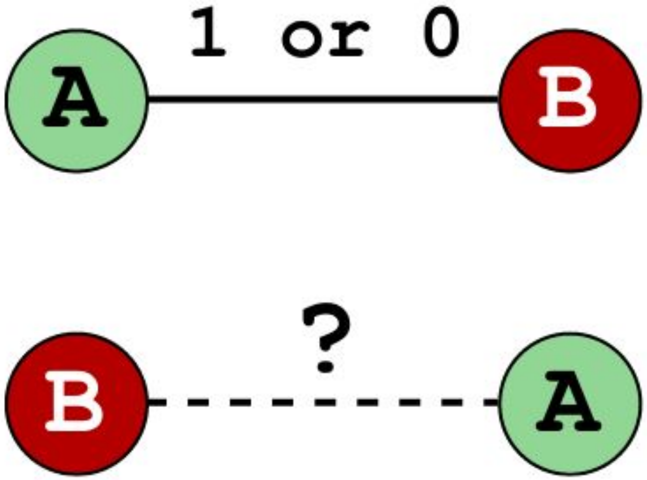


Model Commutativity

Does this order have an impact on model prediction?

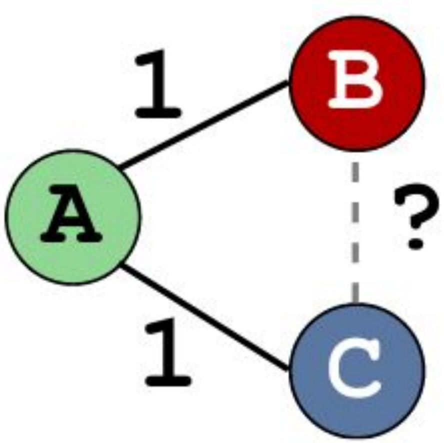
Yes.

Changing headline order changes the prediction **6%** of the time.
Probability shifts on average by 0.06.



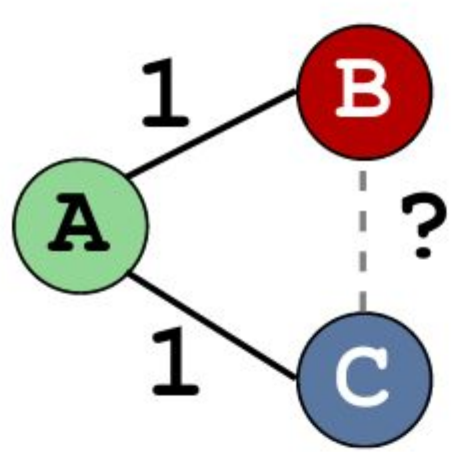
** Tested on Finetune Electra + Time model.*

Model Transitivity



If the model predicts 1 for (A,B), and (A,C), does it predict 1 for (B,C)?

Model Transitivity

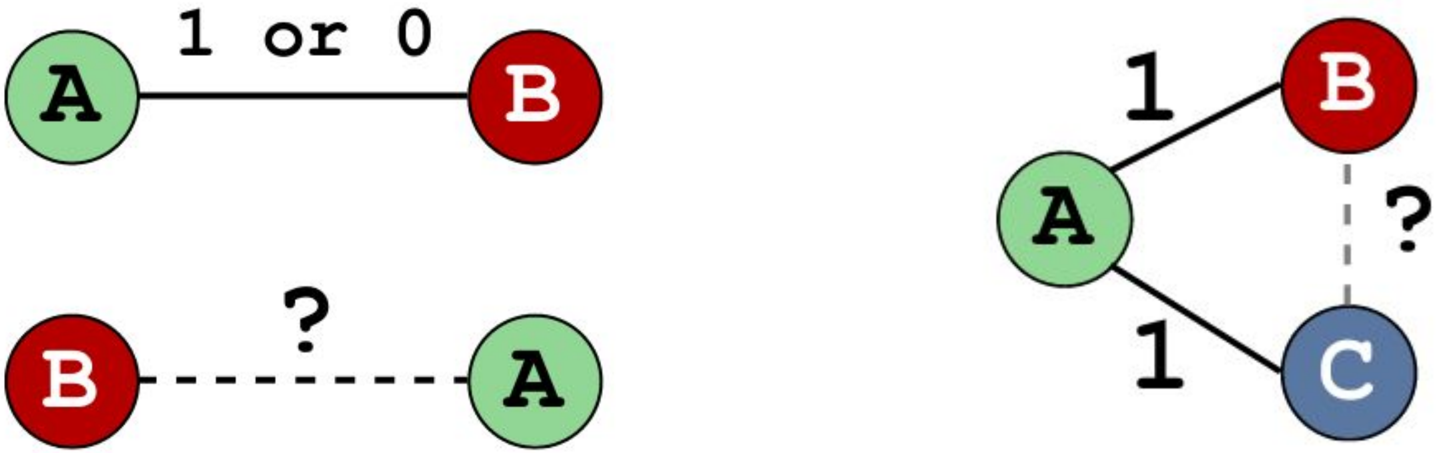


If the model predicts 1 for (A,B), and (A,C), does it predict 1 for (B,C)?

Bad news. The model is consistent only **26.4%** of the time.

** Tested on Finetune Electra + Time model.*

Training Consistent Models?



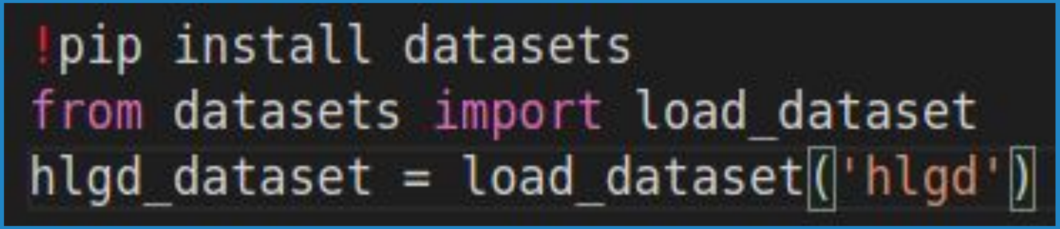
Bid to the listener: Can we train models to be consistent in their prediction when properties are known?

Thanks!

Download HLGD and models:

github.com/tingofurro/headline_grouping

Also available on HuggingFace's *datasets*:



```
!pip install datasets
from datasets import load_dataset
hlgd_dataset = load_dataset('hlgd')
```

Get in touch:

phillab@berkeley.edu

Icon/Avatar credit: *Avatar* | *Flat* from FlatIcon.com

We thank our sponsors!

Bloomberg

amazon web services™

NVIDIA®

Microsoft® Research