WEEK 6

Detecting & Defining Specialized Language in Text

Dr. Santu Karmaker, Dr. Song Wang,
Dr, Mubarak Shah, Adam Bawatneh

Valerie Lopez

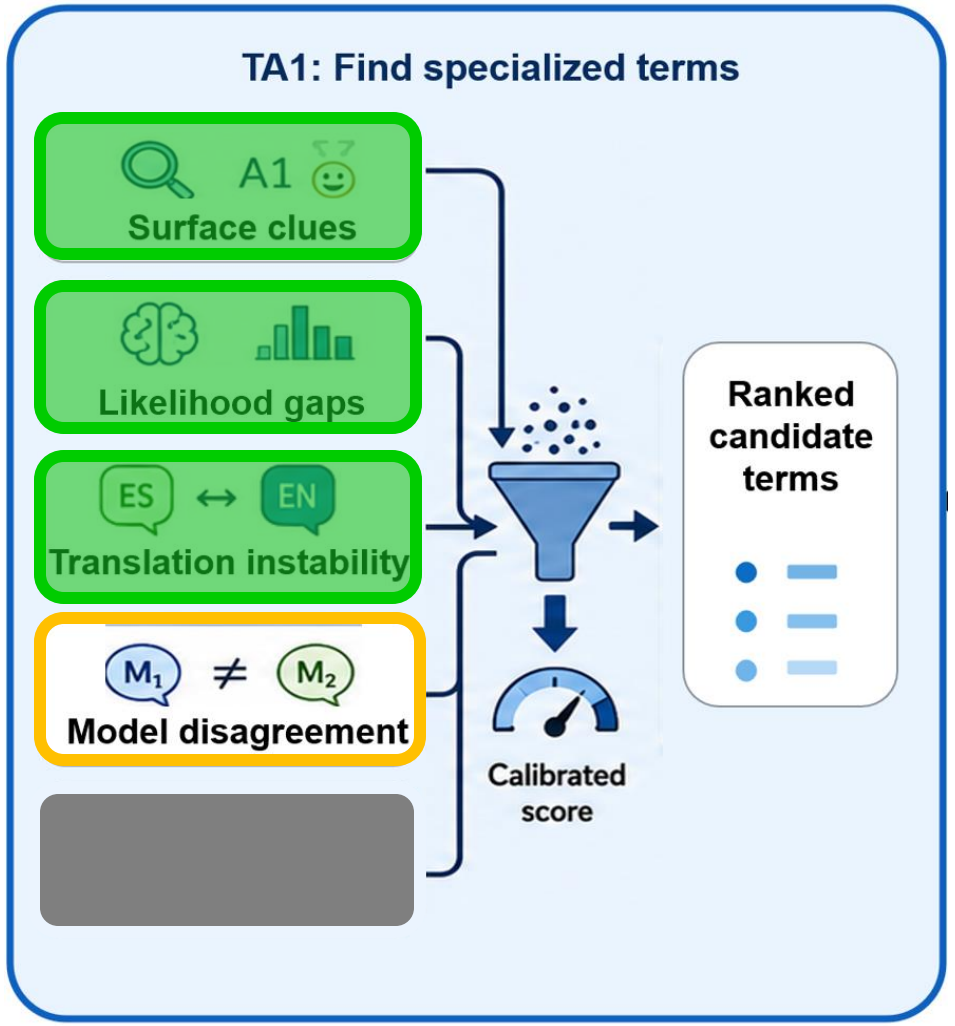


Project Recap

- Identify **special language** from text
- Apply **4** detection **methods**
- **Dataset:** acader/genz-alpha-slangs

This Week:

- Method 4: **Cross Model Agreement**
- Do models agree on definitions?



Method 4: Overview

1. Select a candidate term from given sentence
2. Prompt **3 LLMs** to **define** the candidate
3. If the definitions **disagree strongly**, the candidate is specialized.

Two Approaches: **Semantic Text Similarity vs Natural Language Inference**

Semantic Text Similarity: Overview

1. **Select a candidate term** from given sentence
2. Prompt **3 LLMs** to **define** the candidate
3. Use an STS model to compute **the semantic text similarity** of the definitions
4. For 3 definitions, compute pairwise:
   - Definition 1 vs 2
   - Definition 1 vs 3
   - Definition 2 vs 3
5. Average the pair scores
6. If **average < threshold**, we have **low agreement**. **Conclude specialized!**

STS: Models

Definition Models:

Meta Llama 3.1 8B Instruct
Qwen 2.5 7B Instruct
Mistral 7B Instruct v0.3

STS Model:

STSB RoBERTa Large

- Fine-tuned on Semantic Textual Similarity Benchmark
- Cross encoder that reads 2 texts together
- Returns semantic closeness score

STS Example

Bro, this vibe is straight **mid**.

Llama: Lacking enthusiasm or energy, unimpressive, or uninteresting.

Qwen: Middle, central part.

Mistral: Moderate or average in quality or intensity.

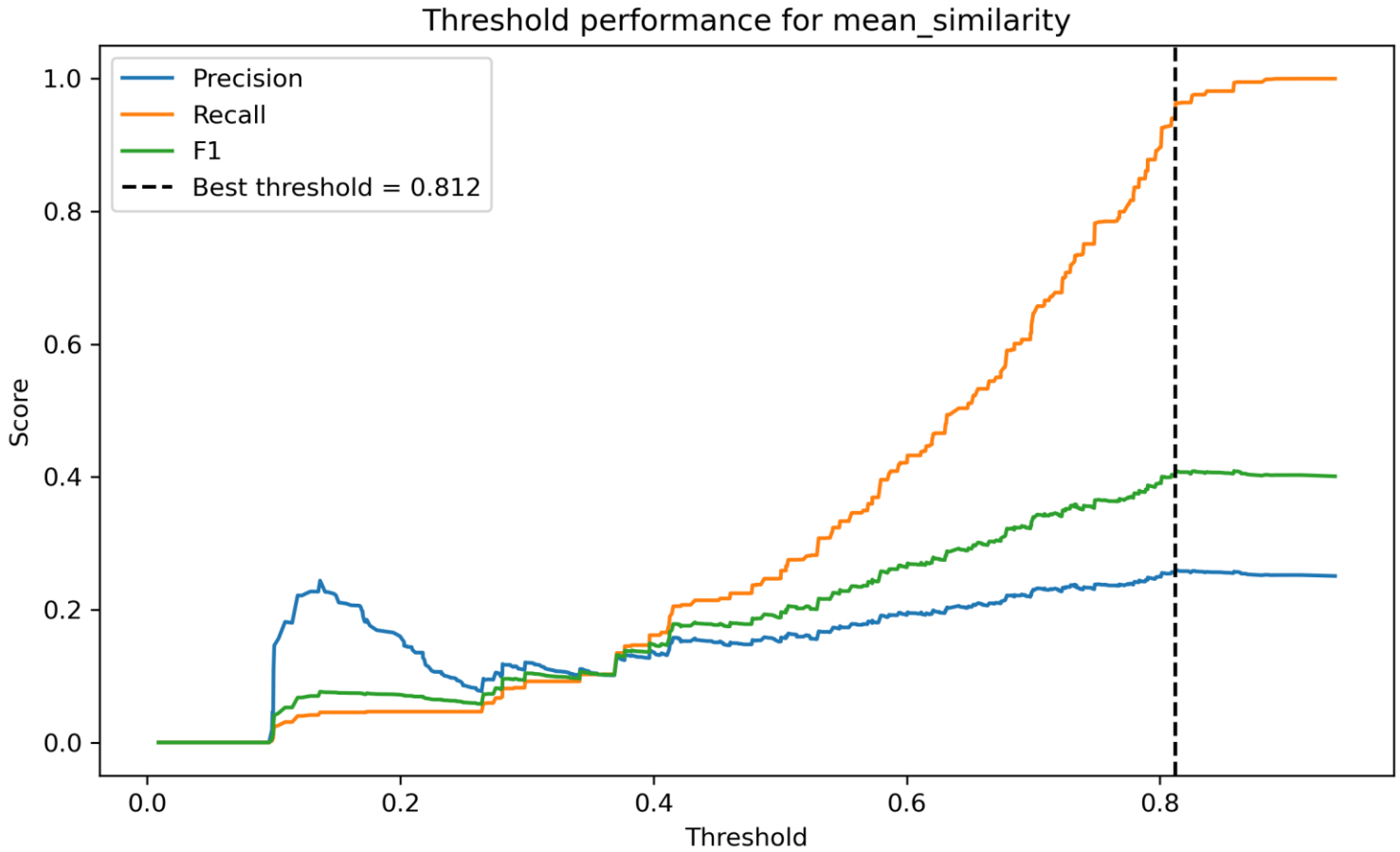
Model Pairing	Similarity Score
Llama-Qwen	**0.01**
Llama-Mistral	**0.27**
Qwen-Mistral	**0.01**

Mean: 0.1

Threshold: 0.812

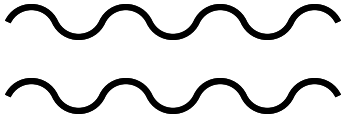
Prediction:
High disagreement.
Candidate is specialized!

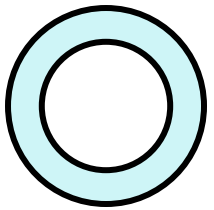
STS Model Results



Precision	Recall	F1
0.26	0.96	0.41

Natural Language Inference





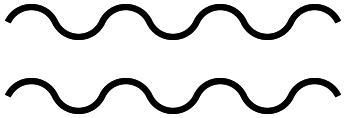
For a pair of definitions (A and B):

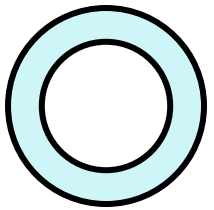
1. **Entailment**: A implies B
2. **Neutral**: A and B are related but there is little entailment nor contradiction
3. **Contradiction**: A and B are logically incompatible

Focus on Entailment only

- Entailment and Contradiction return the same predictions for 99.7% candidates
- Entailment and Contradiction share the same precision, recall, and F1
- Neutral does not provide enough information in either direction

NLI: Overview





1. NLI is directional so we **run twice per pair**
    1. A is the premise, B is the hypothesis
    2. B is the premise, A is the hypothesis
2. **Average these scores**
    1. Average entailment, average neutral, average contradiction
3. **Low entailment** or high contradiction implies **low agreement**.
4. Since we have three pairs, we compare 2 approaches: **majority vote vs average**.

Method 4: NLI Example

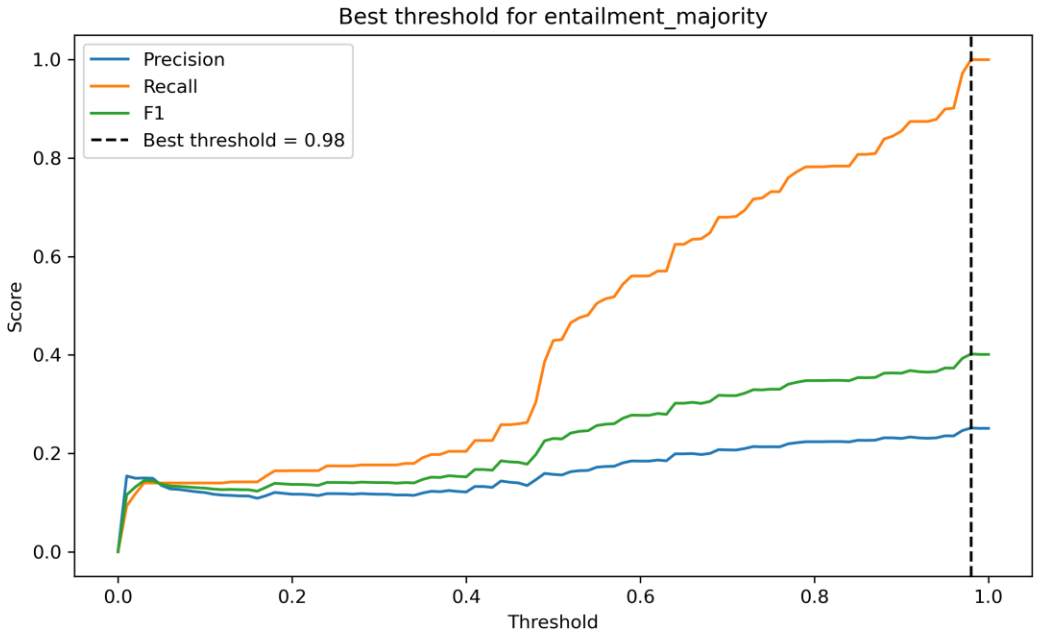
Mate, IRL you fold faster than a Walmart chair

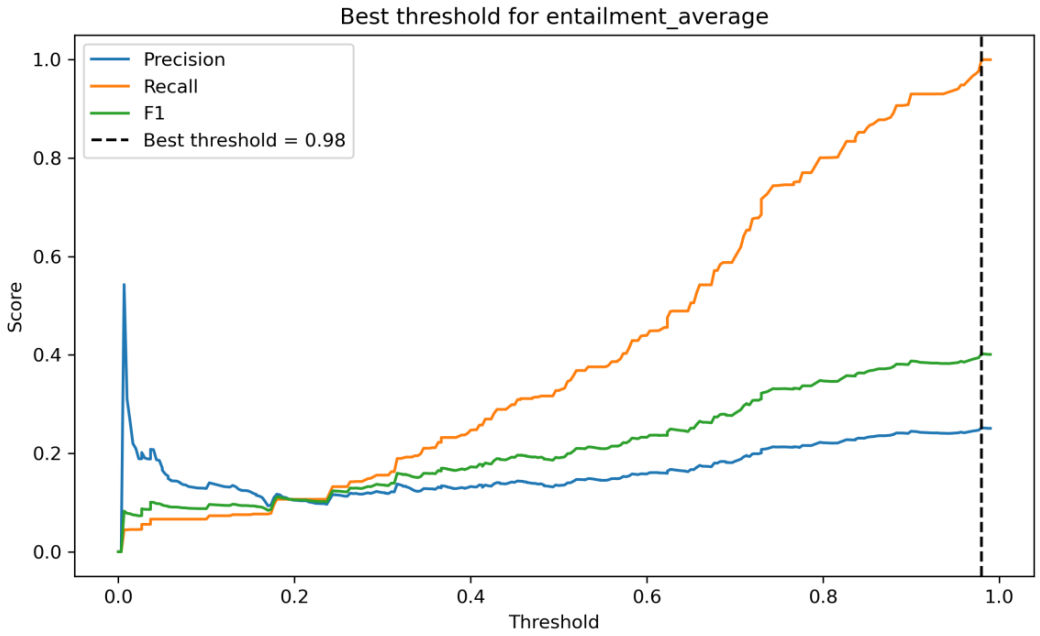
Llama: A close friend or acquaintance, often used in informal or casual online interactions.
Qwen: A term used to address someone familiar or close.
Mistral: Informal term for a friend or acquaintance.

Model Pairing	A to B	B to A	Avg
Llama-Qwen	0.99	0.0	0.49
Llama-Mistral	0.99	0.0	0.49
Qwen-Mistral	0.96	0.99	0.97

Threshold: 0.98
Majority: 3/3
Avg Score: 0.66

NLI: Majority vs Average





	Precision	Recall	F1
Majority Rules	0.252	1	0.402
Average	0.252	1	0.402

STS vs NLI: Example

Fam, stop cappin’

	STS score	NLI (Avg)	STS prediction	NLI Prediction	Is Specialized
fam	0.28	0.02	✓	✓	✓
stop	0.88	0.91	x	✓	x
cappin’	0.54	0.62	✓	✓	✓

	Thresholds
STS	0.81
NLI	0.98

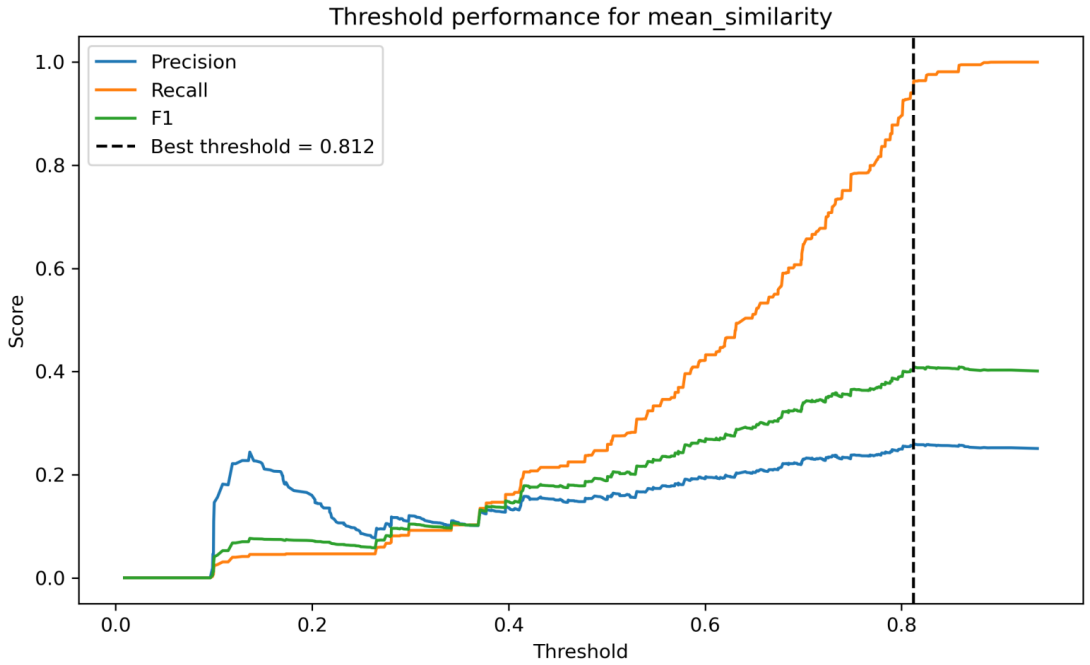
STS vs NLI: Example

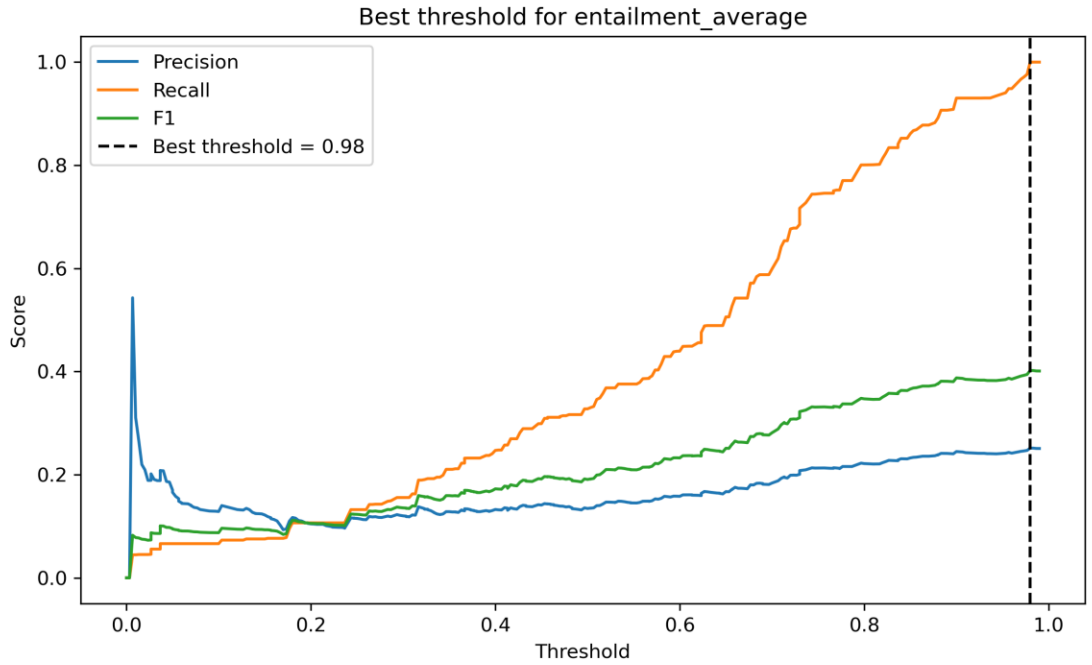
Bro, whole squad waitin’

	STS score	NLI (Avg)	STS prediction	NLI Prediction	Is Specialized
bro	0.70	0.34	✓	✓	✓
whole	0.82	0.63	x	✓	x
squad	0.57	0.15	✓	✓	x
waitin’	0.50	0.62	✓	✓	✓

	Thresholds
STS	0.81
NLI	0.98

STS vs NLI Avg





	Precision	Recall	F1
STS	0.26	0.96	0.41
NLI (Avg)	0.252	1	0.402

Comparing 4 Detection Methods

	Precision	Recall	F1
Surface Clues	0.46	0.49	0.48
Likelihood Gap	0.22	0.55	0.31
Machine Translation (Judge)	**0.49**	0.95	**0.65**
Model Agreement (STS)	0.26	**0.96**	0.41

Improving Current Dataset

Current dataset: acader/genz-alpha-slangs

Mate, IRL you fold faster than a Walmart chair 💀 😂
Homie, you ghosted again 😭 💀 😤 🥹 💀
Dude, pull up or stay sus 💁 💀 🥹
Fam, lowkey flexin’ 💎 🔥 💀 🔥
Bruh, big yikes 😬 💀 😭

- Roughly reviewed ~200 rows
- Noticed that our dataset reuses sentences but rotates **terms of address**
- Double checked with CodeX

Improving Current Dataset

To fix data:

1. We saved the list of these terms
2. For each sentence, we manually added the term to Explanation column
3. Saved the updated dataset

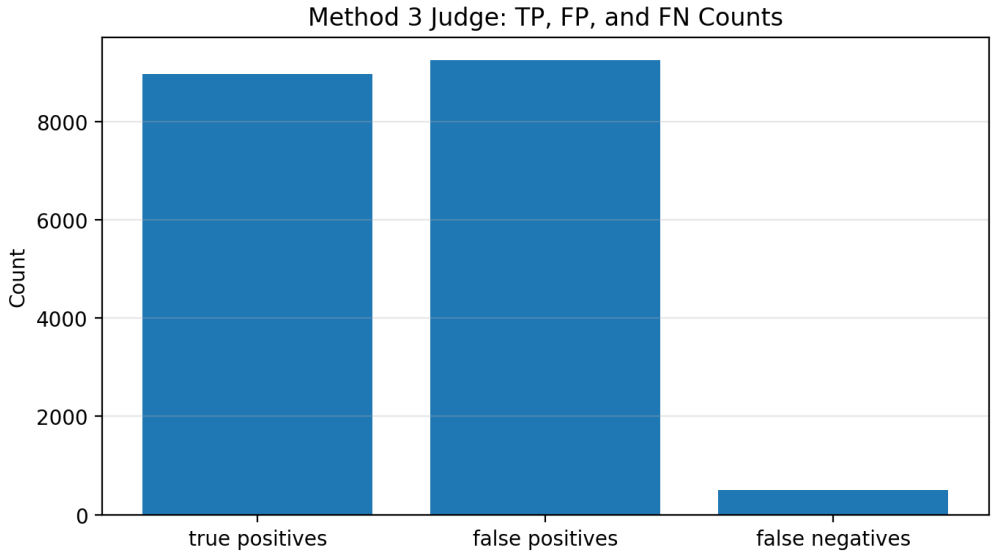
Term	Row count
Mate	645
Homie	640
Bro	587
Bruh	642
Fam	594
Sis	613
Dude	683
Yo	596

Updated Results

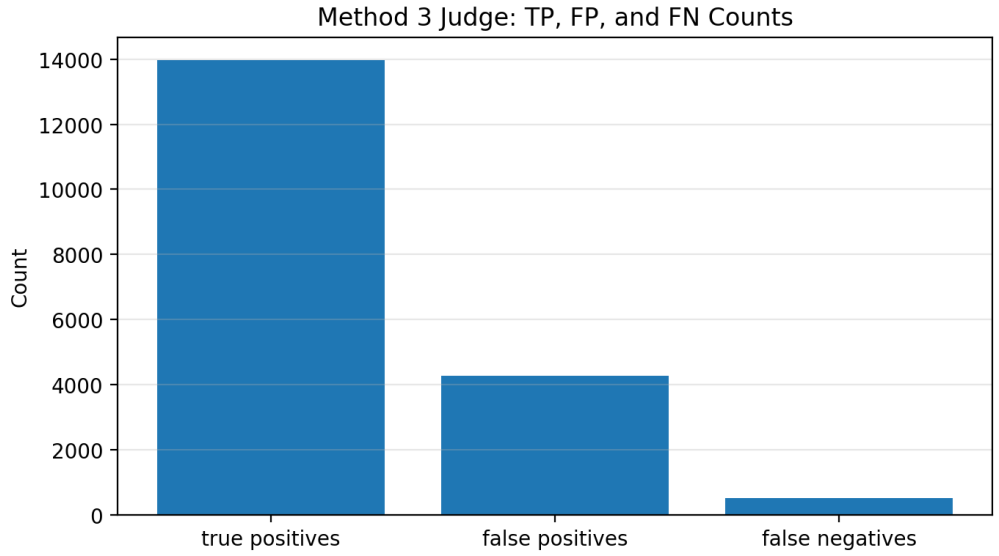
	Precision	Recall	F1	Change in Performance
Surface Clues	0.46	0.29	0.35	Worsened
Likelihood Gap	0.36	0.52	0.43	Improved
Machine Translation (Judge)	**0.77**	**0.97**	**0.85**	Improved
Model Agreement (Semantic)	0.5	**0.97**	0.66	Improved

Next Steps: Building Pipeline

- Machine Translation has the best recall and precision
- Machine Translation is more efficient since it requires only 1 LLM per sentence
- Next Step: **Use MT as the starter method** in the pipeline



Old Dataset



Improved Dataset

Next Steps: Synthesizing Data

- Make synthetic data using glossary type datasets
- Try new domains instead of slang only

LM-Lexicon/Slang
MedReadMe
MACRONYM
SDU@AAAI-22
PARSEME 1.3
AlphaMWE

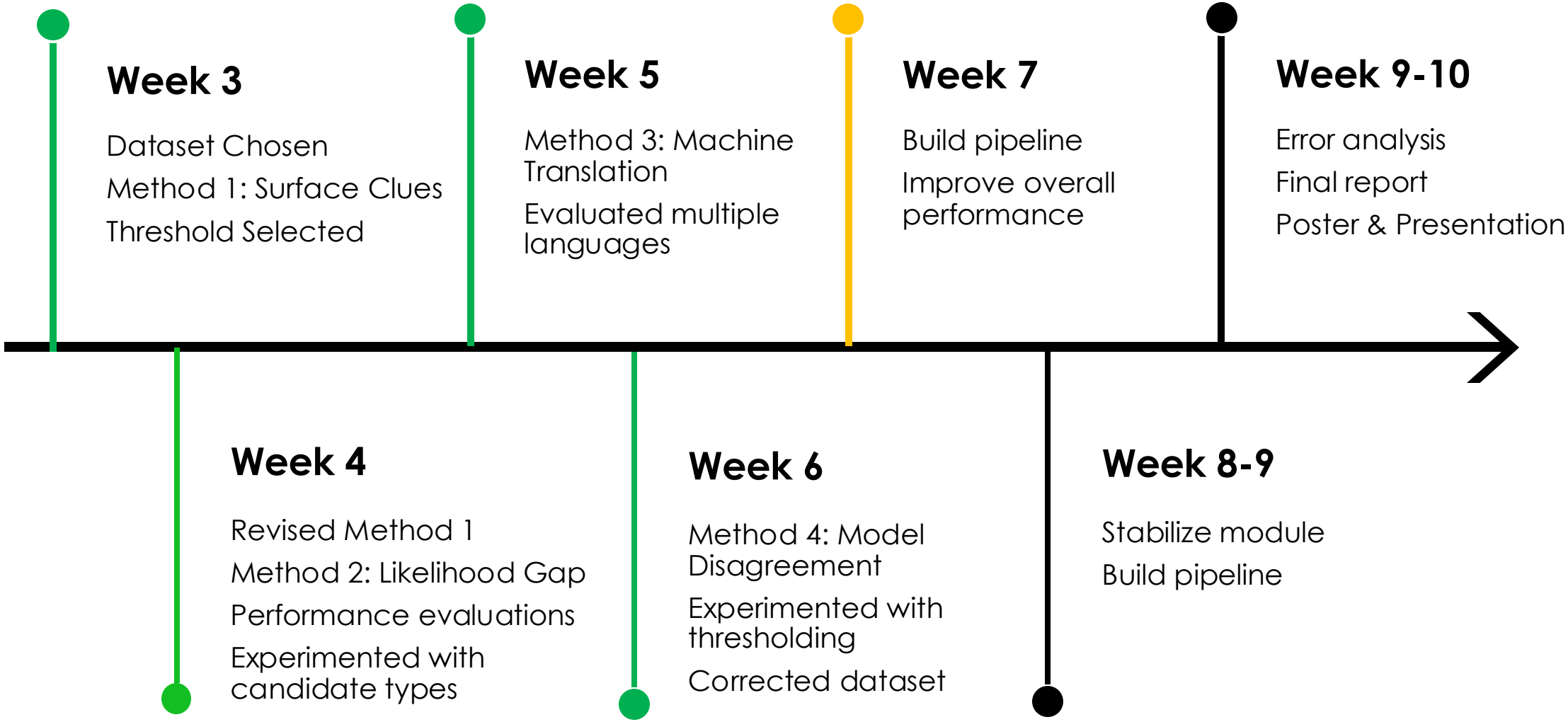
eu		the eu had 15 member states…	eu is the european union, a…	slang	[]
eu		[ey] [maeng], eu need to lie down [d:]	"eu is a synonym for the word for "you""	slang	[]
eu		eu taking you to armageddon with…	european union	slang	[]

Checklist

- ✓ Completed all 4 Detection Methods
- ✓ Fixed current dataset
- ✓ Studied updated performance results

- ❑ Build a pipeline between the 4 methods
- ❑ Move into more specialized language by synthesizing data

Updated Timeline



THANK YOU

Valerie Lopez