

Week 3

Project: Detecting and Defining Specialized Language in Text

VALERIE LOPEZ

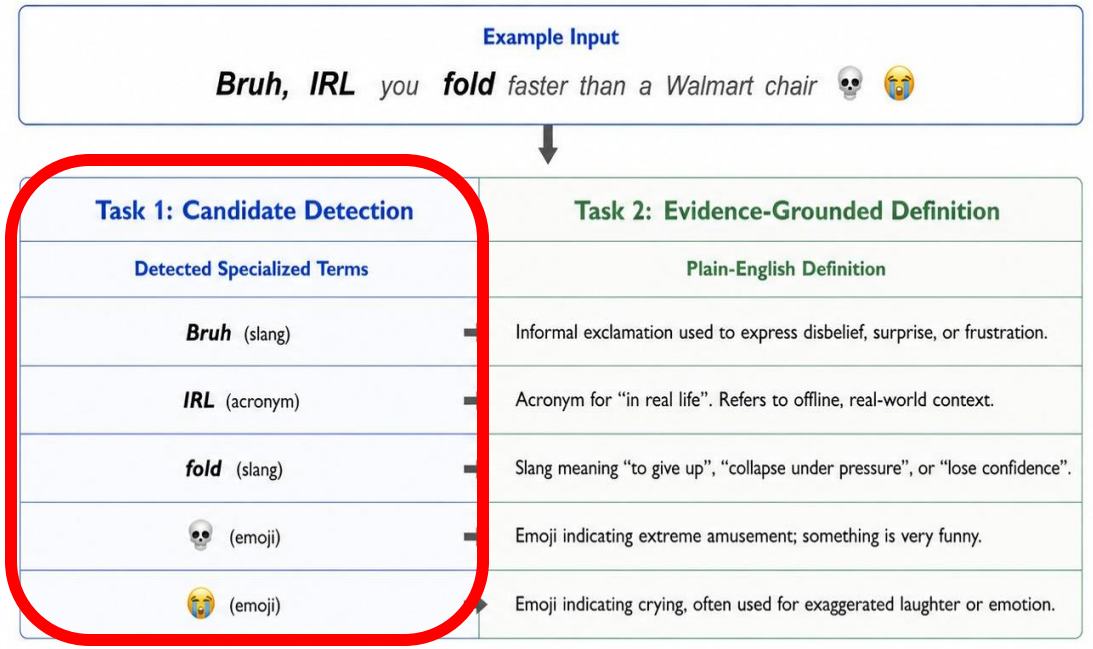
Project Recap

Objective:

- **Identify** and classify **candidates**
  - Include confidence level

Approach:

- Apply multiple **detection filters** to determine how likely a word is to be specialized text



Detection Filters: Overview

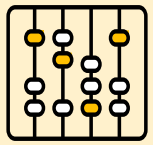
Surface Clues



Do we see alpha-numeric characters, emojis, acronyms?

Based on initial feature **observations**

Language Surprise



How unexpected is this word?

Measures perplexity and token **prediction likelihood**

Machine Translation



Does the machine produce unstable translations?

Instability indicates specialized status

Model Agreement



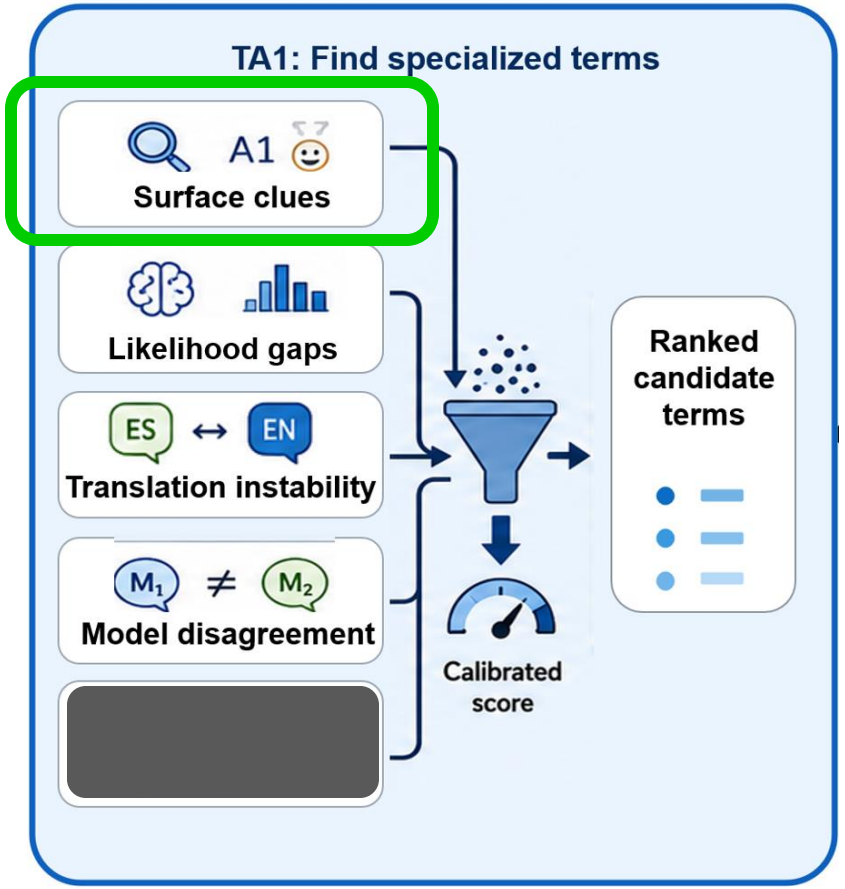
Do our models agree on if the word is a noun, entity, technical term?

Disagreement indicates specialized status

Module 1 Progress

Surface Clues:

- Applying RegEx, ZipF, Emojis **libraries**
- Each candidates receives a **Method 1 score: surface + corpus**
- Determine **thresholds** for specialized terms



Method 1: Surface Clues

Overview:

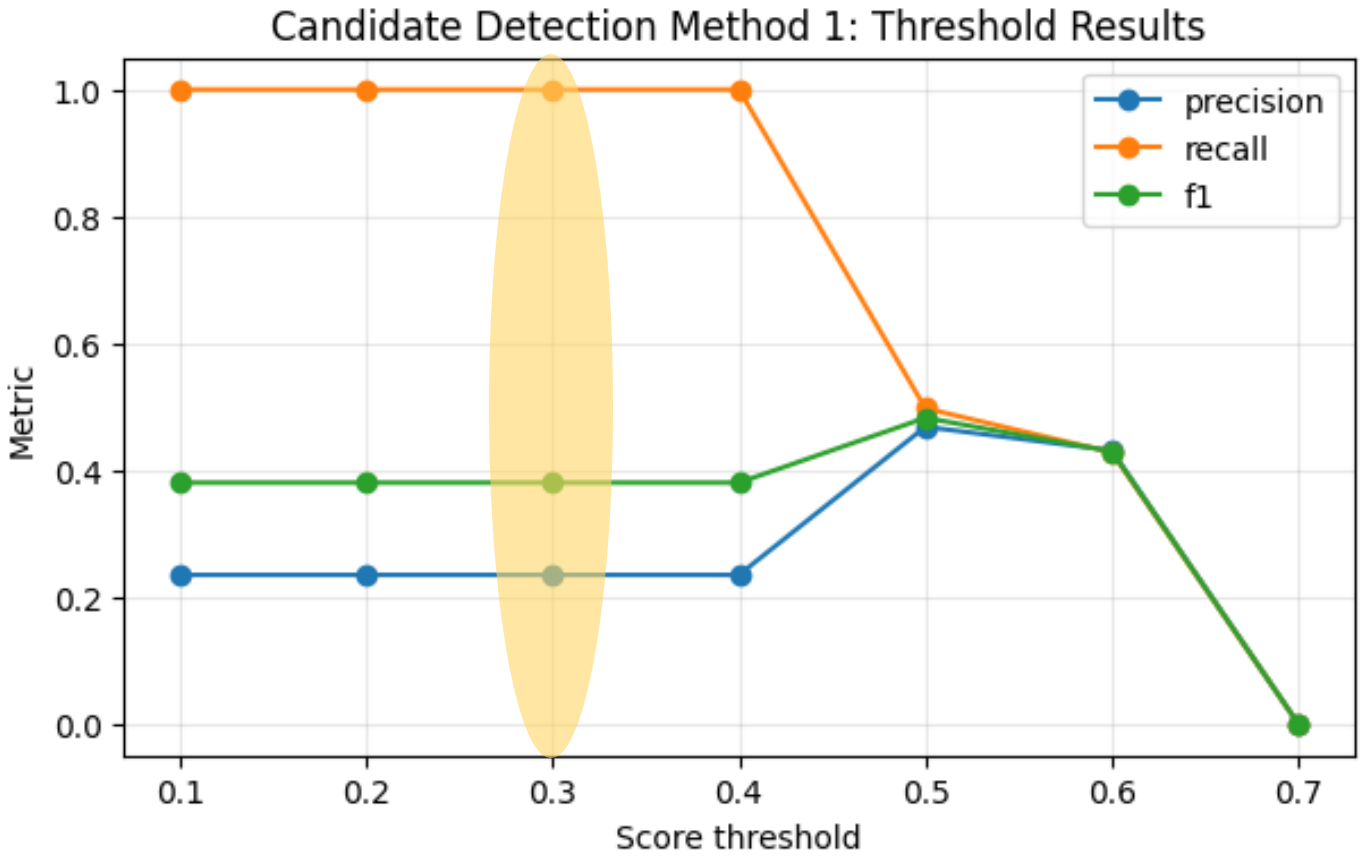
- **Create candidates** based on permutations of words/phrases
- Calculate **surface score**:
  - Frequency, emojis, acronyms, digits, symbols, repeated characters
- Calculate **corpus score**:
  - Specialized terms appear more in examples than translations
- **Experiment** with different **thresholds** to maximize recall

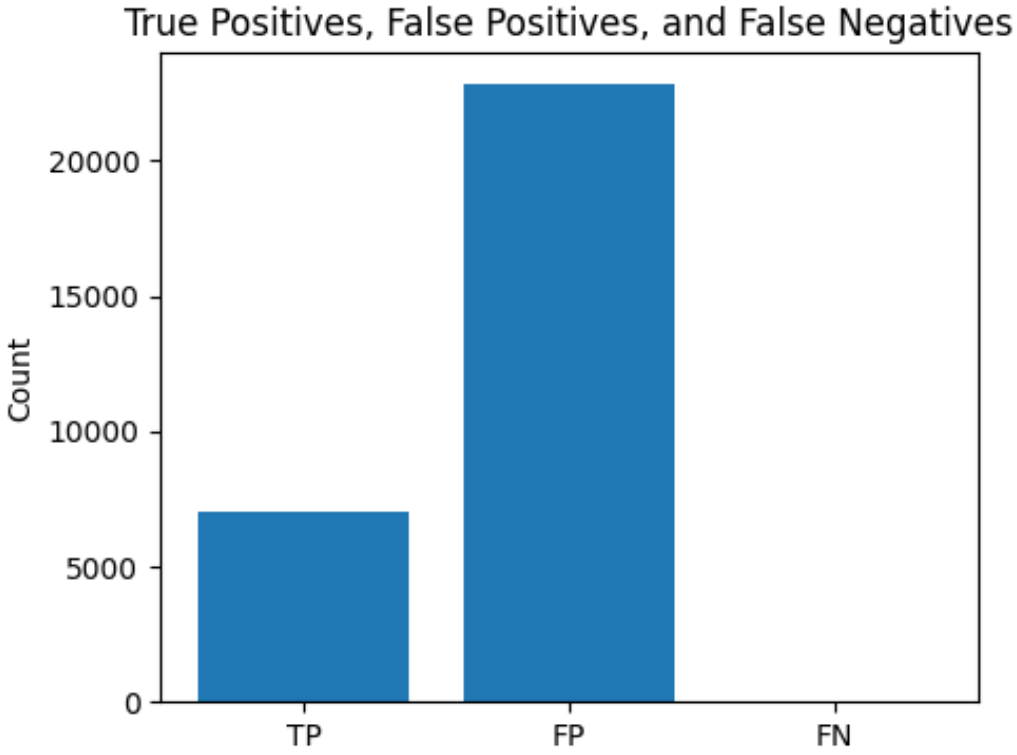
Outputs

Top 5 candidates with highest total scores.

	candidate_norm	surface_score	corpus_score	gold_count	example_sentence
23	ghosted again	0.428571	1.0	0	Bro, you ghosted again 😭💀💎😭
17	flexin	0.428571	1.0	503	Bro, lowkey flexin' 💎🔥🥹🥹
9	cappin	0.428571	1.0	503	Fam, stop cappin', lowkey annoyed 😤🔥😂🥹
34	lowkey	0.428571	1.0	1006	Bro, lowkey flexin' 💎🔥🥹🥹
22	ghosted	0.428571	1.0	482	Bro, you ghosted again 😭💀💎😭

Selecting a Threshold





Threshold: 0.3

Maximize True Positives | Increase recall | Minimize False Negatives | Avoid discarding valid terms too early

Investigating Positives

True positives

	candidate	method1_score	surface_score	corpus_score	sentence	gold_terms
44	ghosted	0.657143	0.428571	1.0	Bro, you ghosted again 😭💀 💎😤	[ghosted]
75	lowkey	0.657143	0.428571	1.0	Bro, lowkey flexin' 💎 🔥 🥹🥹	[lowkey, flexin']
51150	skibidi outta here	0.657143	0.428571	1.0	Bro, skibidi outta here, sigma style 🗿 🔥 😤 🔥	[skibidi outta here]
51019	flexin	0.657143	0.428571	1.0	Dude, lowkey flexin' 💎 🔥 🤷 ⏰	[lowkey, flexin']

False positives

	candidate	method1_score	surface_score	corpus_score	sentence	gold_terms
51090	you ghosted again	0.657143	0.428571	1.0	Mate, you ghosted again 😭💀 💀😂	[ghosted]
46	you ghosted	0.657143	0.428571	1.0	Bro, you ghosted again 😭💀 💎😤	[ghosted]
47	ghosted again	0.657143	0.428571	1.0	Bro, you ghosted again 😭💀 💎😤	[ghosted]
48	you ghosted again	0.657143	0.428571	1.0	Bro, you ghosted again 😭💀 💎😤	[ghosted]

FP rate is high due to candidate search without meaning.

Method 2: Language Surprise

XLM-RoBERTA Model:

- Encoder finetuned on 100+ languages
- “Fill-Mask” useful for likelihood gap scoring
  - Great for domain specific problems

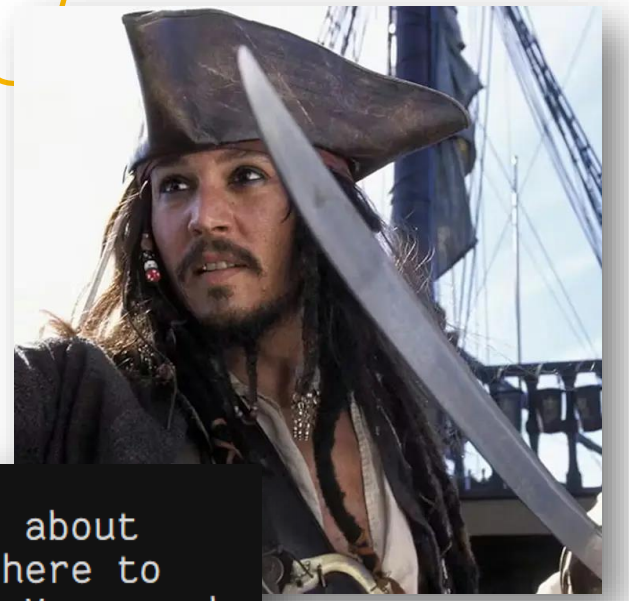
This Week’s Focus:

- Implement RoBERTa model for token prediction likelihood



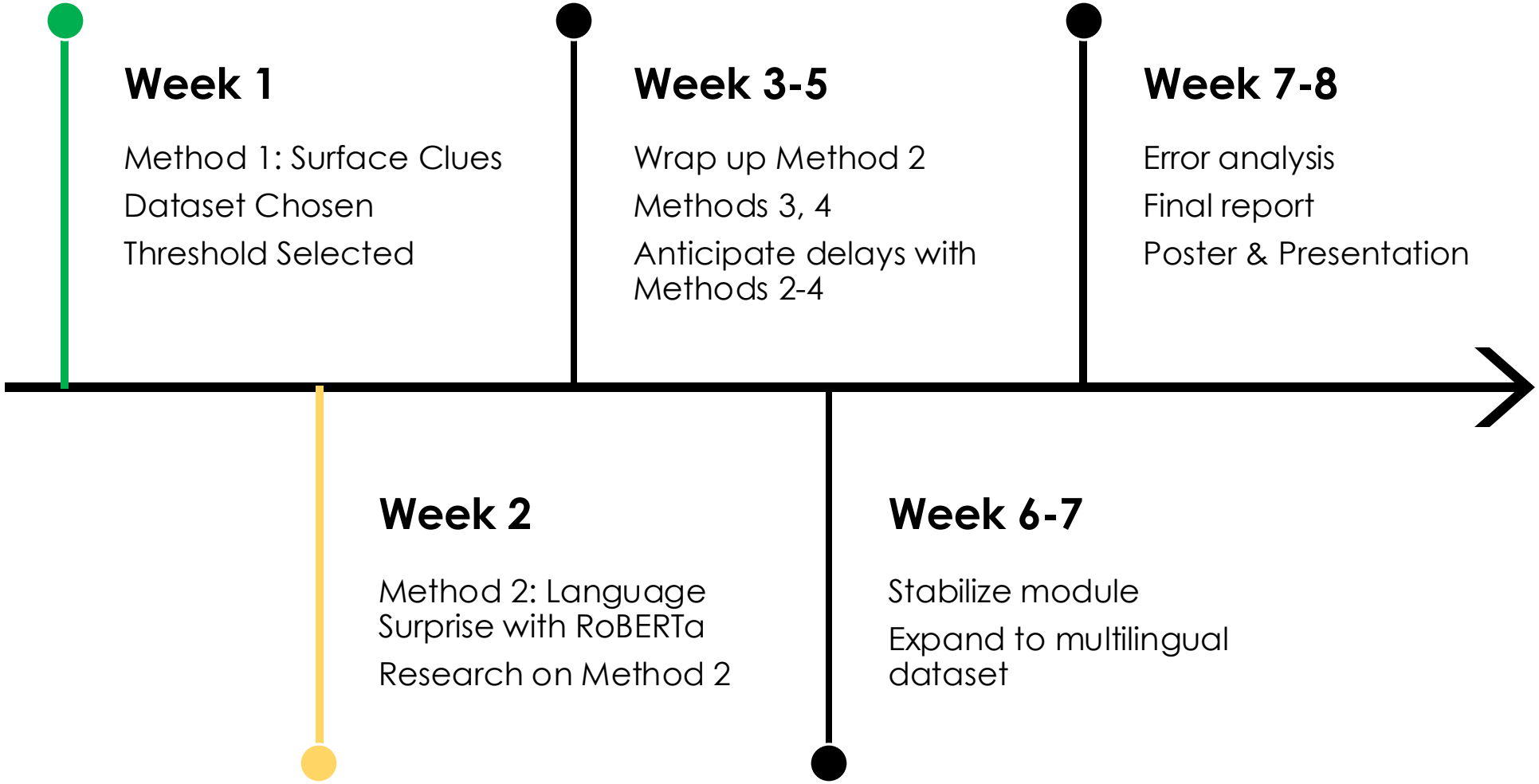
Meta Llama: Understanding tools for later methods

- **Meta Llama 3.1-8B Instruct Model**
- Ran **inference** on the model
- Gained **insight on related skills** necessary for implementing future candidate detection methods
  - Loading models from **HuggingFace**
  - Allocating GPU on **Newton**
  - **Prompting** model and receiving responses



```
Model loaded successfully
{'role': 'assistant', 'content': "Yer lookin' fer a bit o' info about
 meself, eh? Well, matey, I be a swashbucklin' pirate chatbot, here to
 provide ye with treasure troves o' knowledge and witty banter! Me name be
 Captain Chatbeard, and I've spent me fair share o' years sailin' the
 seven seas o' the internet, plunderin' the riches o' knowledge and
 bringin' 'em back to ye, me hearties! So hoist the sails and set course
 fer a treasure-filled conversation, savvy?"}
```

Updated Timeline



Thank you

Any questions?