Automatic Information Extraction from Unstructured Text – *practices and case studies*

SHUBHANSHU MISHRA

ISCHOOL AT ILLINOIS

@THESHUBHANSHU

Structured v/s Unstructured Text

Structured Text

Persons: Rafa→wiki(Rafael_Nadal), Fed→wiki(Roger Federer)
Event: USOpen→wiki(US Open 2017)
Time: May 2nd, 2017
Sentiment: Support towards Rafa

- Rarely available
- Unique representation
- Easy to parse and load in database
- Easy to answer questions on
- Easy to search

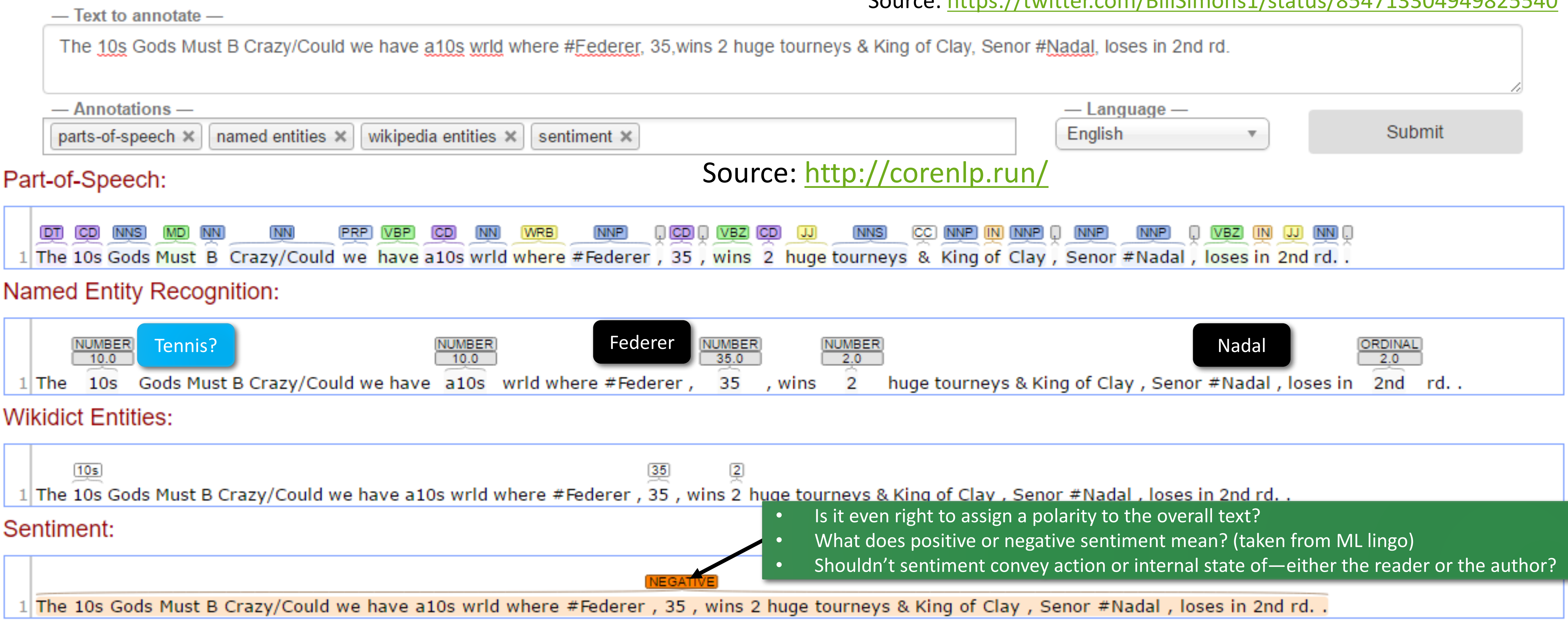
Un-Structured Text

Rafa and Fed take on each other on May 2nd #USOpen #GoRafa.
Timestamp: April 31st, 2017

- Abundantly available
- Multiple representations
- Easy to parse if data is proper English (e.g. newswire corpus)
- Harder to parse when the language used is different

Information extraction issues

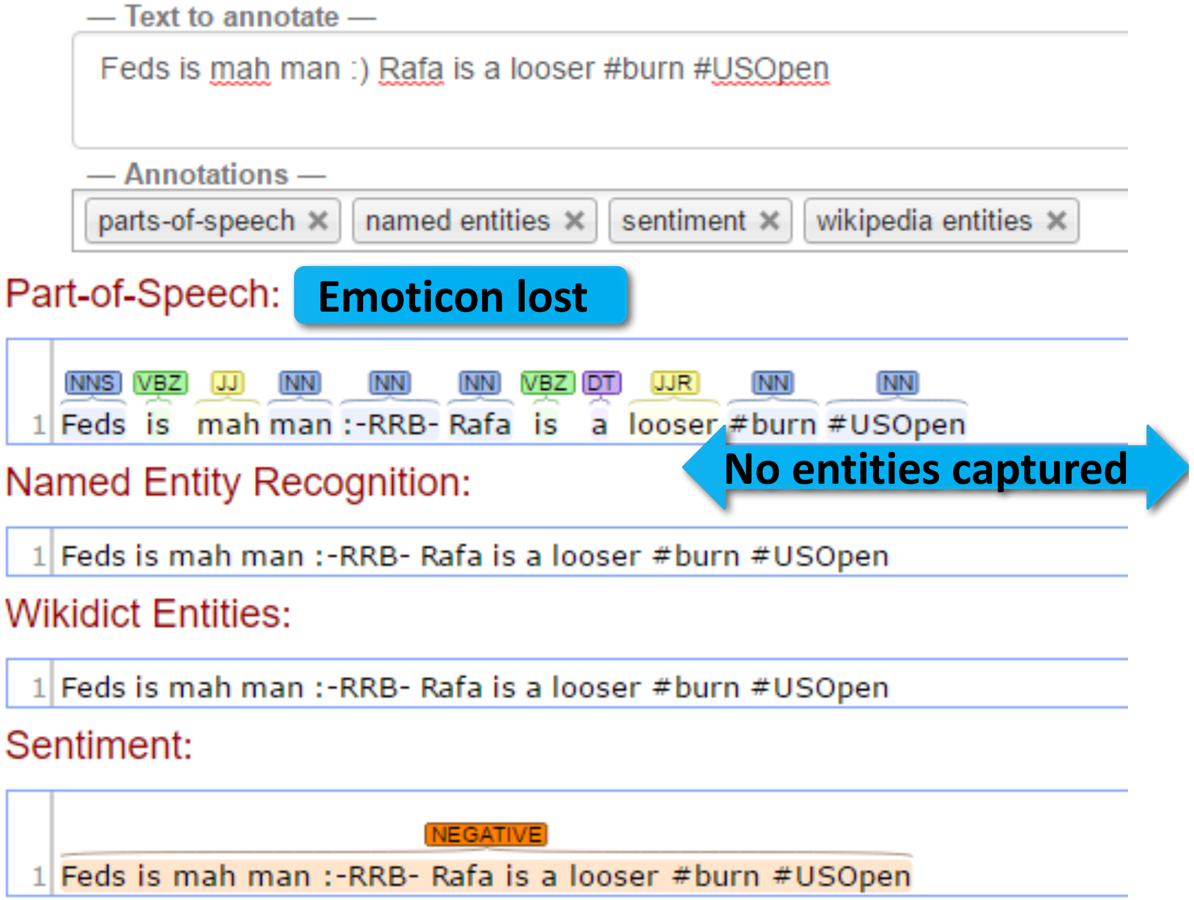
Source: https://twitter.com/BillSimons1/status/854713304949825540

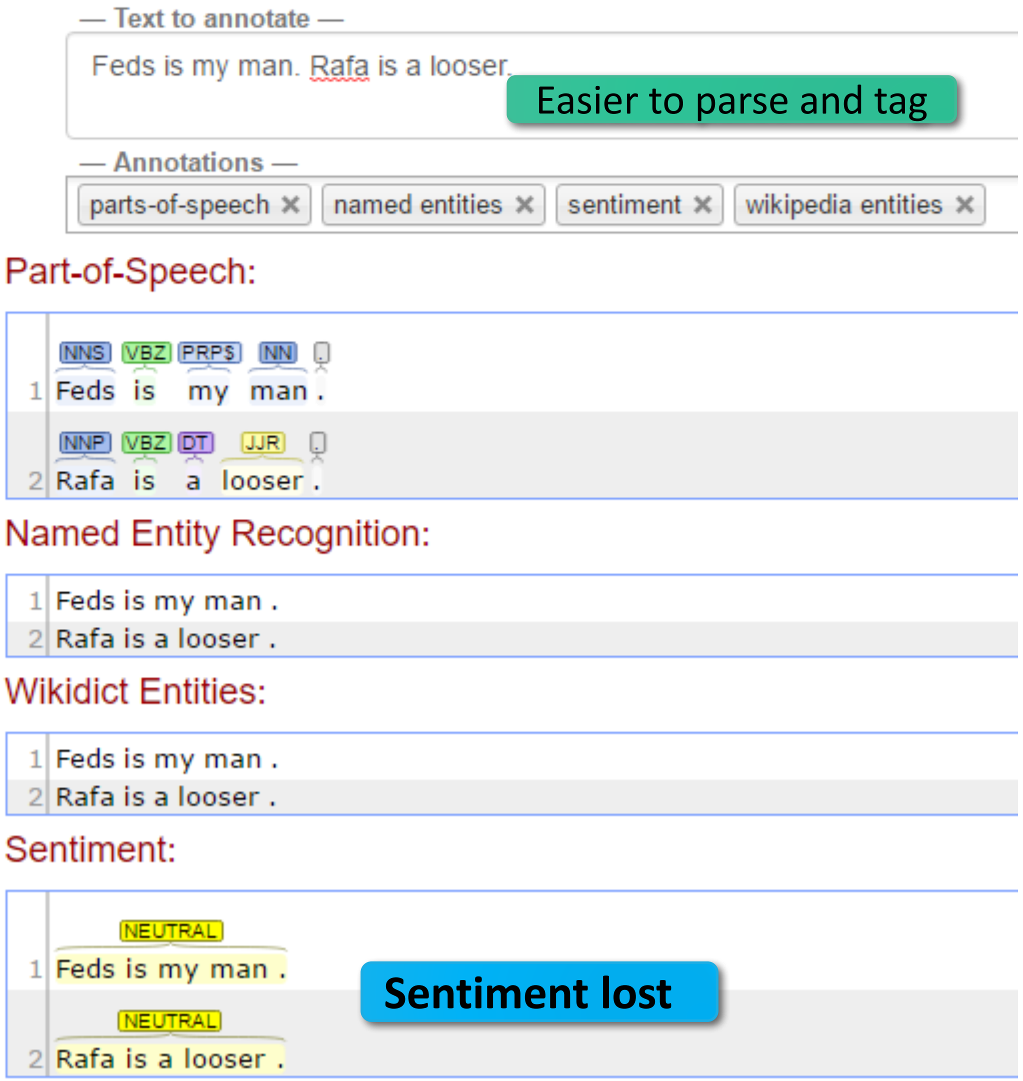


Source: http://corenlp.run/

- Is it even right to assign a polarity to the overall text?
- What does positive or negative sentiment mean? (taken from ML lingo)
- Shouldn’t sentiment convey action or internal state of—either the reader or the author?

Traditional NLP tools break on Social Media Data





Source: http://corenlp.run/

*"Our field is the **domain science of language technology**; it's **not** about the best method of machine learning—the central issue remains the **domain problems**. The domain problems **will not go away**."*

--- ***Christopher Manning*** *(ACL President 2015)*

Source: ***http://mitp.nautil.us/article/170/last-words-computational-linguistics-and-deep-learning***

Solutions for NLP in Social Media Text

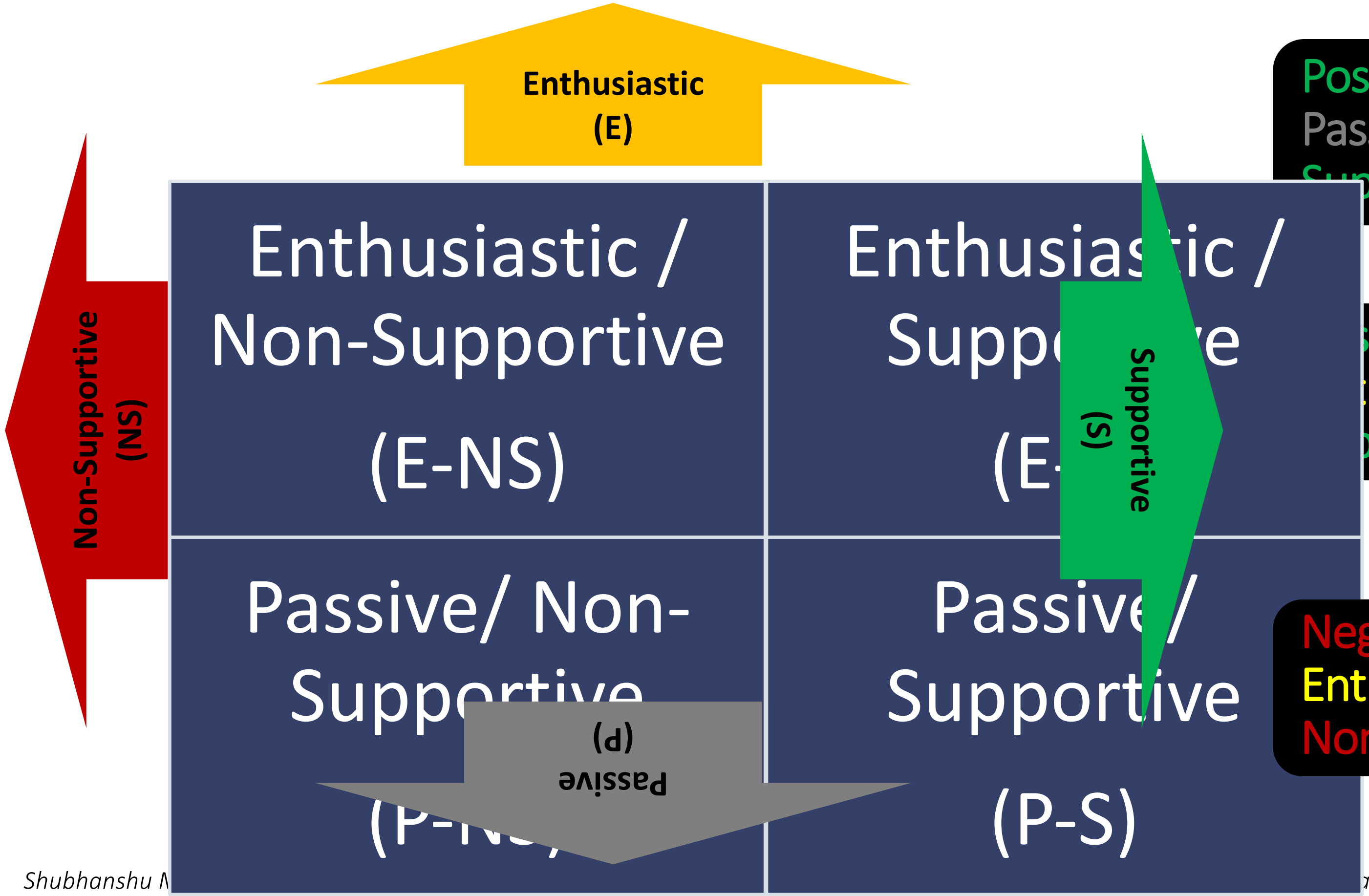
Sentiment Analysis:

- Define ***actionable labels*** for sentiment to capture if the author is:
  - ***Enthusiastic or passive*** in their expression
  - ***Supportive or non-supportive*** of the topic
  - Make task specific labels – don't just stick to positive or negative
- Utilize the ***context of the text*** for making prediction:
  - Who is the ***user***? – number of followers, friends, statuses, etc.
  - What ***metadata*** is associated with the tweet? – retweets, favorites, etc.

Named Entity Recognition and Part of Speech Tagging:

- Allow models to:
  - Change with ***time***
  - Adapt to your testing data
- This can be achieve via:
  - Updating the ***vocabulary*** of the model
  - Using ***online or semi-supervised learning*** models to learn from unlabeled and newer data

Actionable labels:
Enthusiasm and Support



Positive → Passive & Supportive

"New LGBT Research Study on same sex weddings [link]"

sitive → thusiastic & pportive

"All the best to the retired players suffering from CTE. Spread the word so we can make the game safer."

Negative → Enthusiastic & Non-Supportive

"Just watched cyberbully-- it's annoying. Why would she kill herself? It's not worth it. Life is shit so deal with it :P"

Shubhanshu M … alternative sentiment classification for social movements on social media. In Proceedings of the 2014 ACM conference on Web science (WebSci '14). ACM, New York, NY, USA, 261-262. DOI: http://dx.doi.org/10.1145/2615569.2615667

Actionable labels:
Tuition-Free Higher Education Policy Debate on Facebook

Opinion on subject

- ***For:*** Tax money used for a good long-term investment, finally.
- ***Against:*** Free? It's not free-someone's gotta pay.

Tone of the conversation

- ***Civil:*** Nothing is free Mr. President.
- ***Non-Civil:*** NOTHING IS FREE!!!!!!! WAKE UP!!

Relatedness to the topic

- ***On-Topic:*** I hope they pay for it with cuts in wasteful spending, not more taxes.
- ***Off-Topic:*** There's ALWAYS a catch, just like with "You can keep your insurance"

Collier, D.A., Shubhanshu, M., Houston, D., Hensley, B., & Hartlep, N.D. (In Press). Americans "support" the idea of tuition-free college: An exploration of sentiment and political identity suggests otherwise. *Journal of Further and Higher Education.*

Use context
User attributes: Gender, Ethnicity, etc.

Gender

- Extracted via US SSN data
- Can also use existing image based classifiers

Ethnicity

- Predicted based on full name via Textmap tool. http://www.textmap.com/ethnicity/

Race

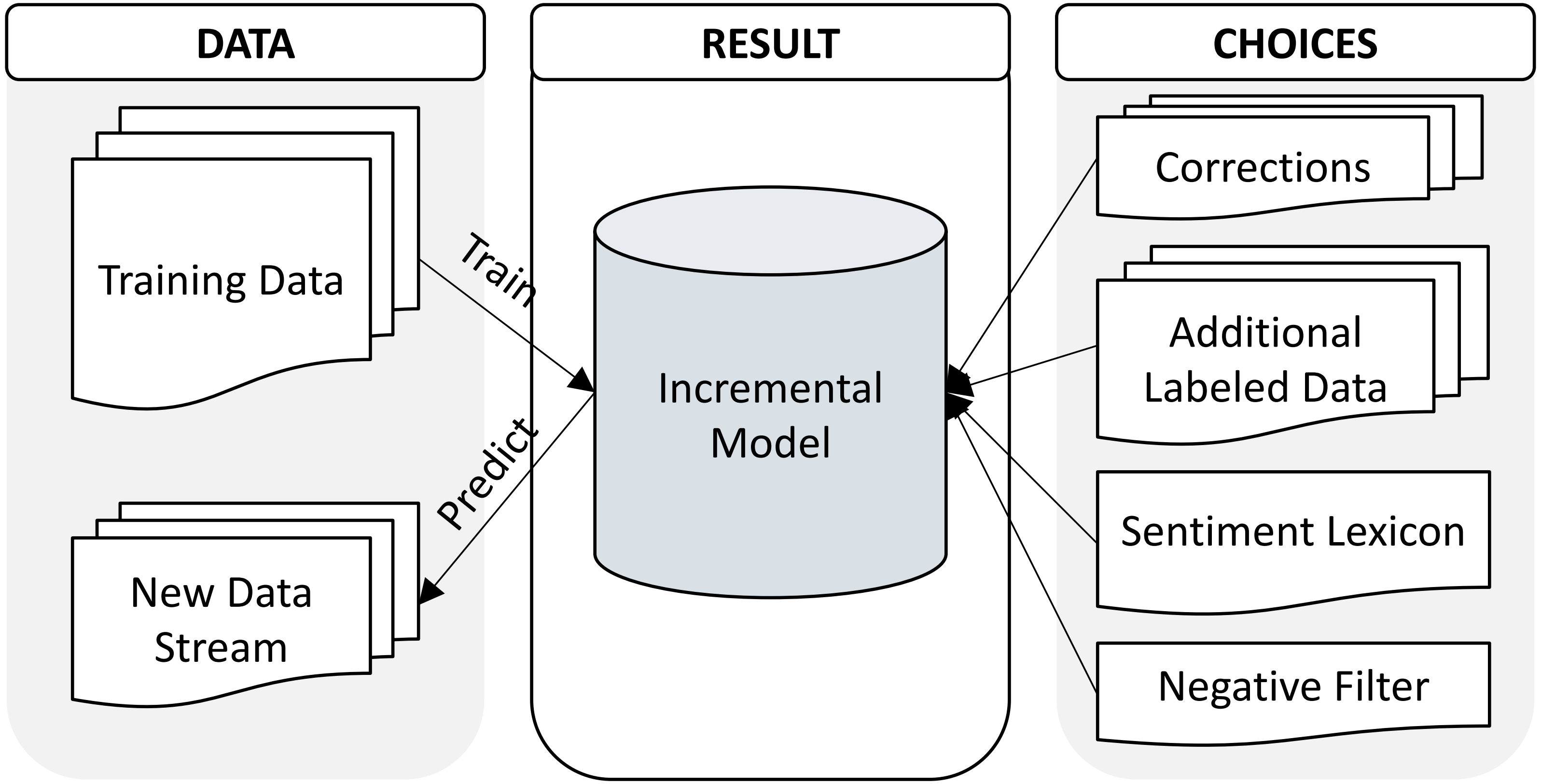
- Manually annotated using profile inspection

Political Leaning

- Manually annotated using profile inspection (if available)

Collier, D.A., Shubhanshu, M., Houston, D., Hensley, B., & Hartlep, N.D. (In Press). Americans "support" the idea of tuition-free college: An exploration of sentiment and political identity suggests otherwise. *Journal of Further and Higher Education.*

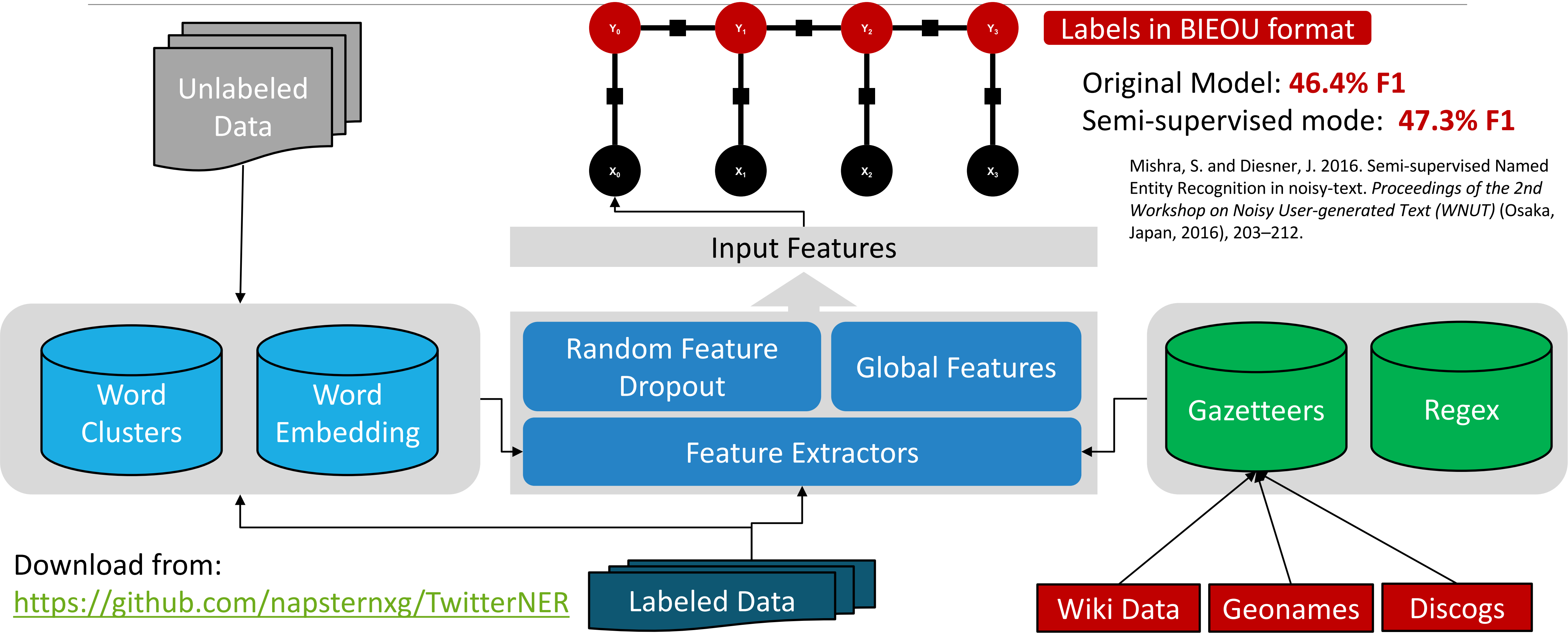
Adaptive models (Sentiment): *Vocabulary Update & Online Learning*



Download tool from:
https://github.com/uiuc-ischool-scanr/SAIL

Shubhanshu Mishra, Jana Diesner, Jason Byrne, and Elizabeth Surbeck. 2015. Sentiment Analysis with Incremental Human-in-the-Loop Learning and Lexical Resource Customization. In Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15). ACM, New York, NY, USA, 323-325. DOI: http://dx.doi.org/10.1145/2700171.2791022

Adaptive models (NER):
Vocabulary Update & Semi-supervised Learning



Original Model: **46.4% F1**
Semi-supervised mode: **47.3% F1**

Mishra, S. and Diesner, J. 2016. Semi-supervised Named Entity Recognition in noisy-text. *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)* (Osaka, Japan, 2016), 203–212.

Download from:
https://github.com/napsternxg/TwitterNER

Open challenges

Not enough labeled data for social media IE tasks

Not enough pre-trained models

Rapidly evolving contents

Privacy and TOS issues – data available today might be missing tomorrow

Key Takeaways

Don’t apply existing tools blindly

Understand your labels and their relation with the data

Use the context

Allow models to evolve (be aware if they don’t)

Questions ?

Contact:

Shubhanshu Mishra

http://shubhanshu.com

@TheShubhanshu