Hierarchical classification for Multilingual Language Identification and Named Entity Recognition

Saatvik Shah **Vaibhav Jain**
Anshul Mittal **Sarthak Jain**
Shubham Tripathi **Jatin Verma**
Rajesh Kumar

Malaviya National Institute of Technology, Jaipur

December 04, 2015

Problem Statement

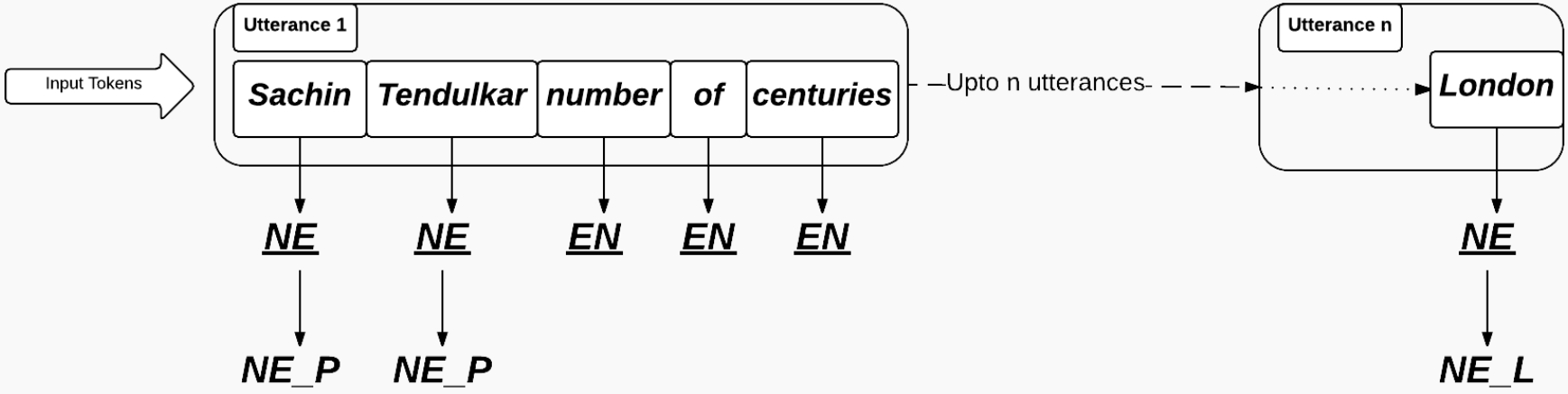
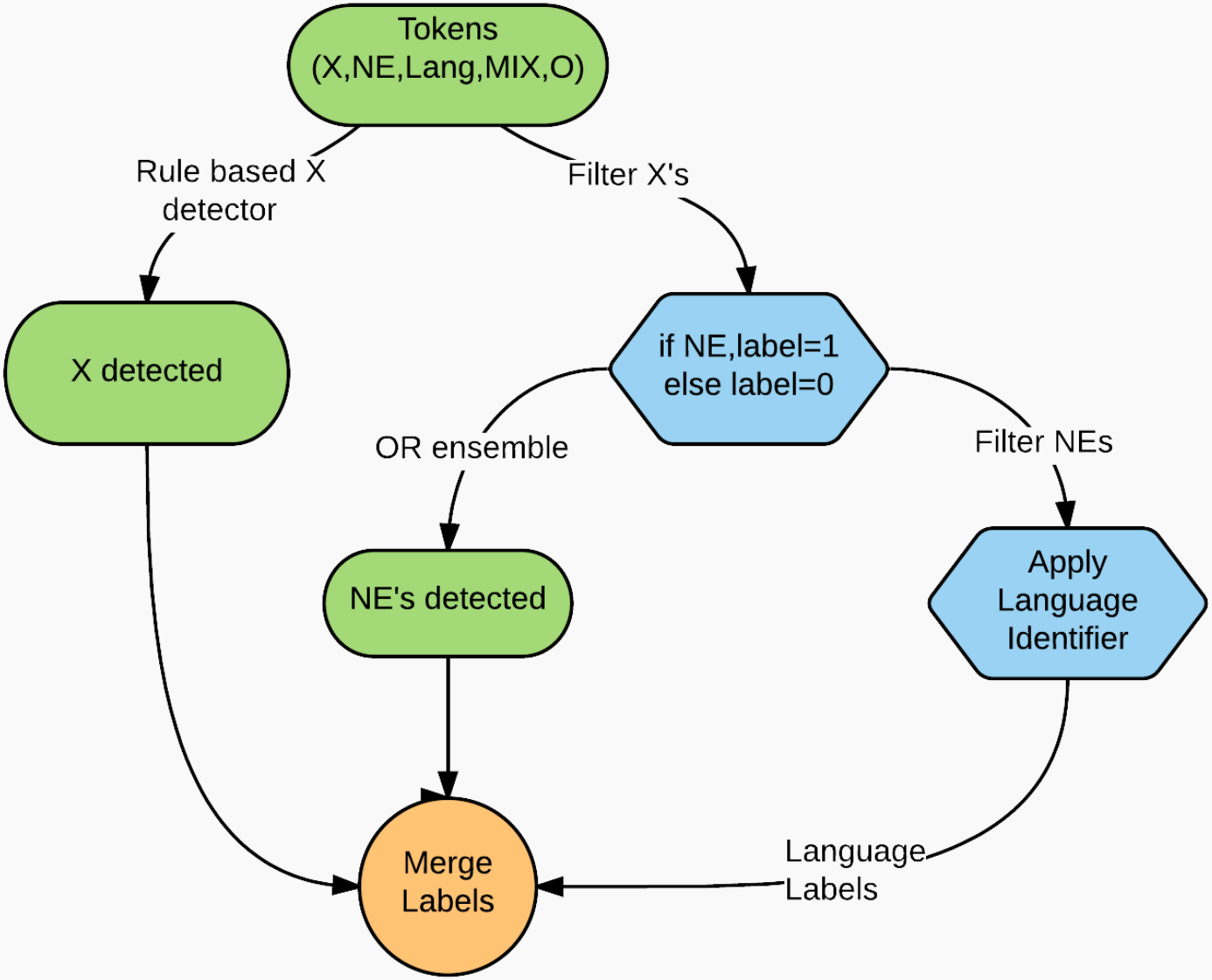


Figure: Task Description

Challenges

- Small dataset
- NE's not subclassified in training data
- 9 languages!!

Hierarchical Classification Workflow



Let's Begin!

Sample Utterance

kohli ki consistency !! IndiAn3 team lineup #wow

Punctuation Recognition

Sample Utterance

kohli ki consistency **!!** IndiAn3 team lineup **#wow**

1. Regex based cleanup to mark tokens such as
   - ***http://*** or ***abc@def.xyz*** for Web URLs
   - **:))**, >**:(**, **:-P** for Emoticons
2. Gazetted List of **Acronyms**

Feature Set[1]

Sample Utterance - {X}

kohli ki consistency **IndiAn3** team lineup

- Word Context - *(ki, consistency, **IndiAn3**, team, lineup)*
- Character level n-gram, 2,3,4,5-gram
  - 3-gram - *(ind),(ndi),(diA),(iAn),(An3)*
- Relative Position - *BEGIN/END Tag*
- Word Normalization - *AaaaAa0*
- Composition Features - *WordDigit*
- POS Tags

and a few more...

Converted to a **Bag Of Words** representation

Named Entity Recognition[2, 3]

kohli ki consistency **IndiAn3** team lineup

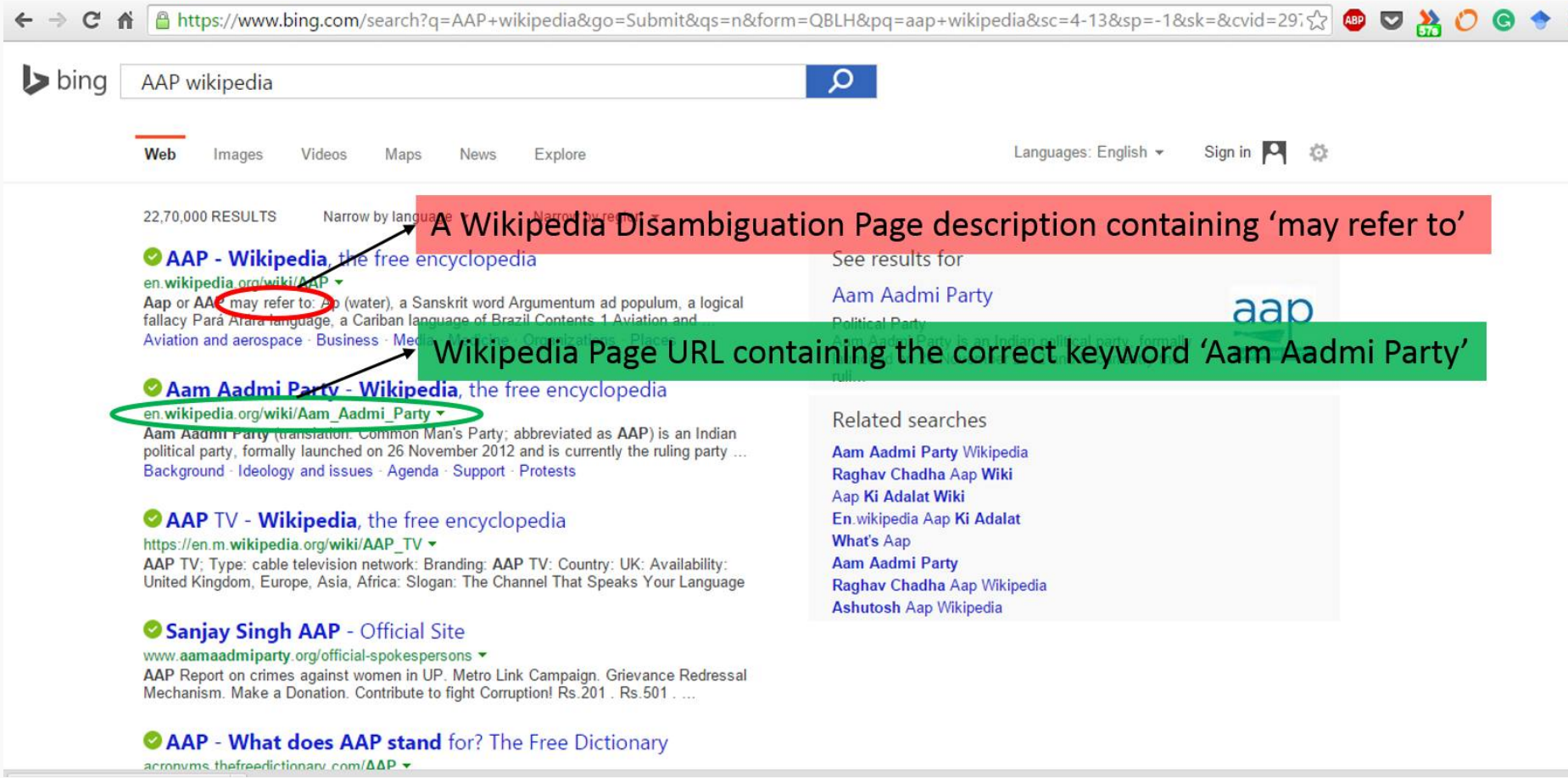
Supervised NE classification

- Binary classification of words : **NE** or **Non-NE**.
- Linear SVM, Logistic Regression and Random forests
- **Logical OR** ensemble for final prediction.

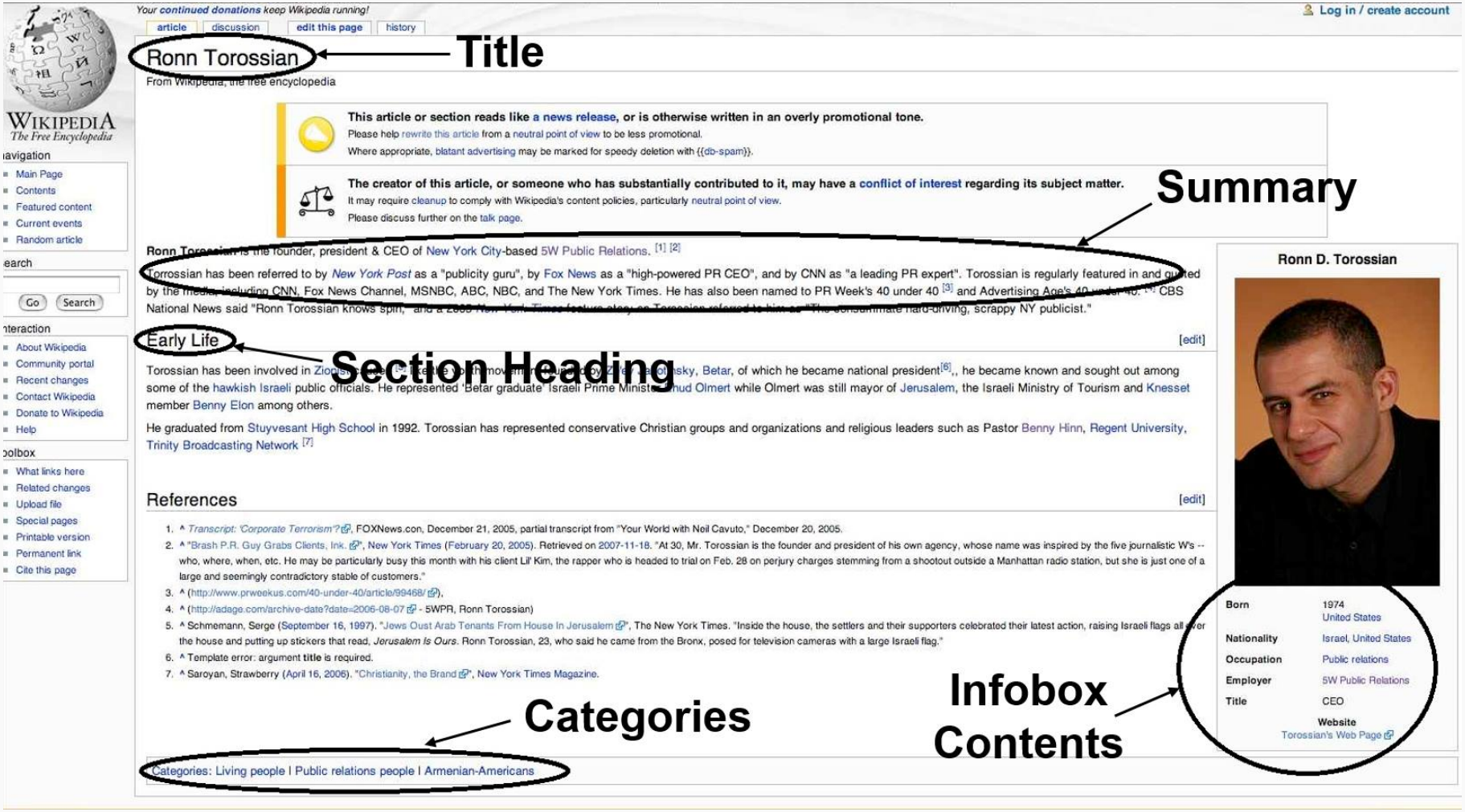
Unsupervised NE subclassification

- Extract Wikipedia Page
- If **Disambiguation**: Search *'$query wikipedia'* on Bing
- Extract most relevant non-Disambiguation
- Apply text keyword scoring

Unsupervised approach - I[6]



Unsupervised approach - II



Language Identification[1, 4, 5]

Sample Utterance - {X,NE}

ki consistency team lineup

- We implemented Linear-Kernel Support Vector Machines
- Important features, on cross validating,
  - Word N-gram, 2,3,4,5-gram
  - Local Knowledge
  - Part of Speech Tags
  - Composition Features

Merged Labels

Sample Utterance

kohli ki consistency !! IndiAn3 team lineup #wow

Labelled Utterance

NE_P hi en X NE_P en en X

Done!

Results

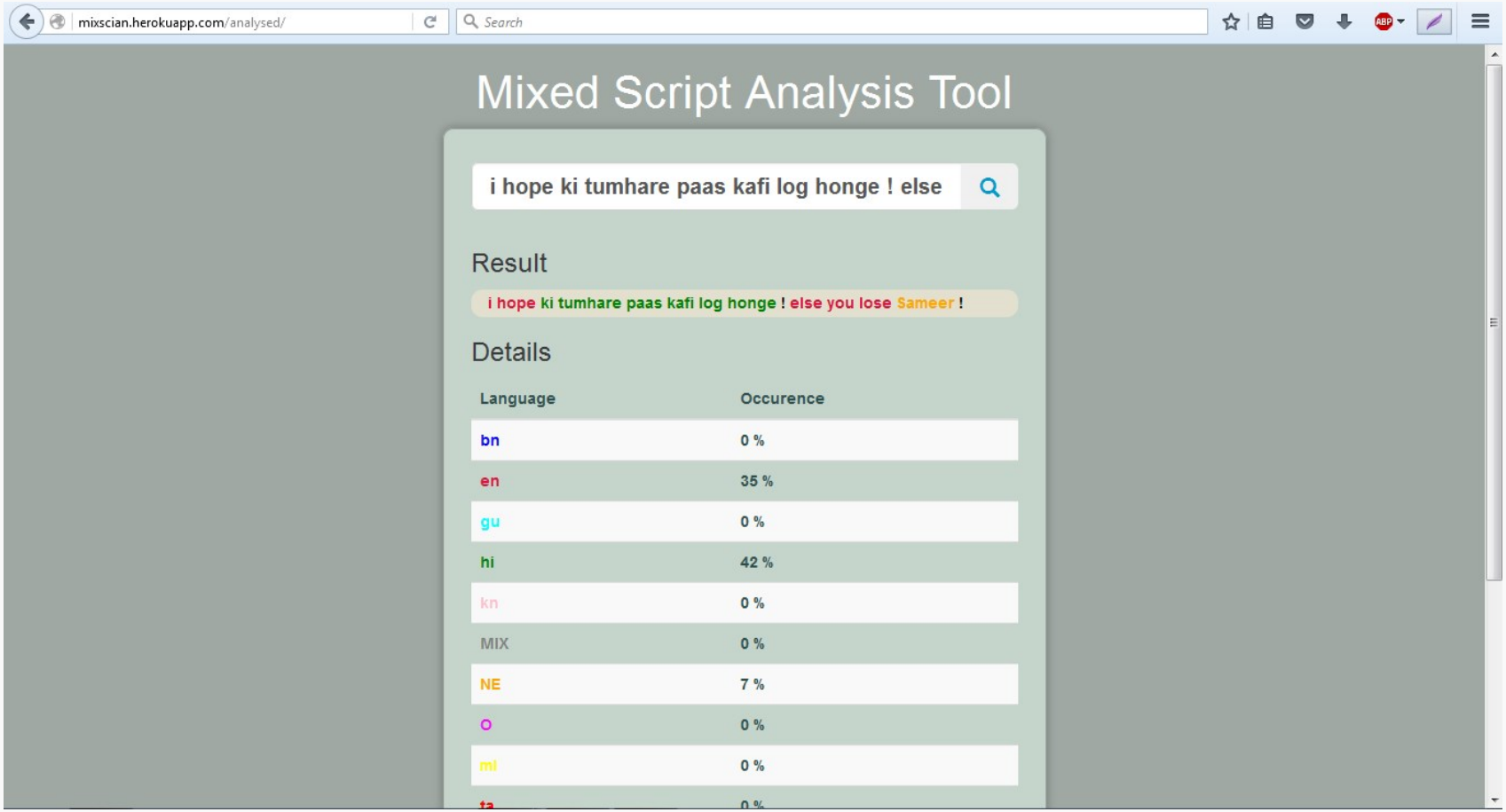
P, R, F-S	Bn	En	Gu	Hi	Kn	Ml	Mr	Ta	Te	X	NE	NE_L	NE_P
P	0.878	0.958	0.097	0.817	0.575	0.394	0.705	0.937	0.431	0.961	0.368	0.722	0.2121
R	0.966	0.848	0.5	0.74	0.829	0.752	0.79	0.708	0.687	0.966	0.528	0.124	0.25
F-S	0.838	0.9	0.163	0.776	0.679	0.517	0.745	0.806	0.529	0.964	0.433	0.214	0.229

Table: Strict Precision (P), Recall (R), F-scores (F-S) for languages and NEs.

COMMENTS:

1. **Weigthed F-measure** = 0.808
2. Gujarati and Hindi are morphologically quite similar.
3. Indian names commonly confused with Hindi

Implementation



- Web Application at **https://mixscian.herokuapp.com**
- Source code at **https://github.com/saatvikshah1994/hline**

Conclusion

1. What worked:
   - Splitting the classification problem into a hierarchy
   - Bing + Wikipedia for highly acccurate NE subclassification
   - LinearSVM
2. What didn't work:
   - IRSLTM
   - LIGA
   - Home built Spell Corrector
   - Stanford NER Tagger
3. What will be done:
   - More Data!!
   - More Gazetted Lists
   - Word Embeddings as a feature for NER
   - Adding a Language Group level to Hierarchy

Acknowledgments

The authors would like to thank the reviewers and organizers.
We also like to thank **ACM SIGIR** for the traval grant provided.

Questions???

[1] Gupta, D. K., Kumar, S., & Ekbal *A. Machine Learning Approach for Language Identification & Transliteration: Shared Task Report of IITP-TS.*. Centre for Digital Music, 2012.

[2] Abhinaya N., Neethu John, Dr. M. Anand Kumar, Dr. K.P. Soman *Amrita @ FIRE-2014: Named Entity Recognition for Indian Languages*. 2014.

[3] Dubey, S., Goel, B., Prabhakar, D. K., & Pal S.. *ISM@ FIRE-2014: Named Entity Recognition Indian Languages*.

[4] King, B., & Abney, S. P. *Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods*. In HLT-NAACL (pp. 1110-1119). 2013.

[5] A. Das and B. Gamback. *Code-Mixing in Social Media Text:The Last Language Identification Frontier? Traitement Automatique des Langues (TAL): Special Issue on Social Networks and NLP , TAL Volume 54 no 3/2013, Pages 41-64*.

[6] Nothman, J., Ringland, N., Radford, W., Murphy, T., & Curran, J. R. *Learning multilingual named entity recognition from Wikipedia*. Artificial Intelligence, 194, 151-175. 2013.