# Building intuitive chat bots via language switching pattern mining

**Sai Kiran Peketi**

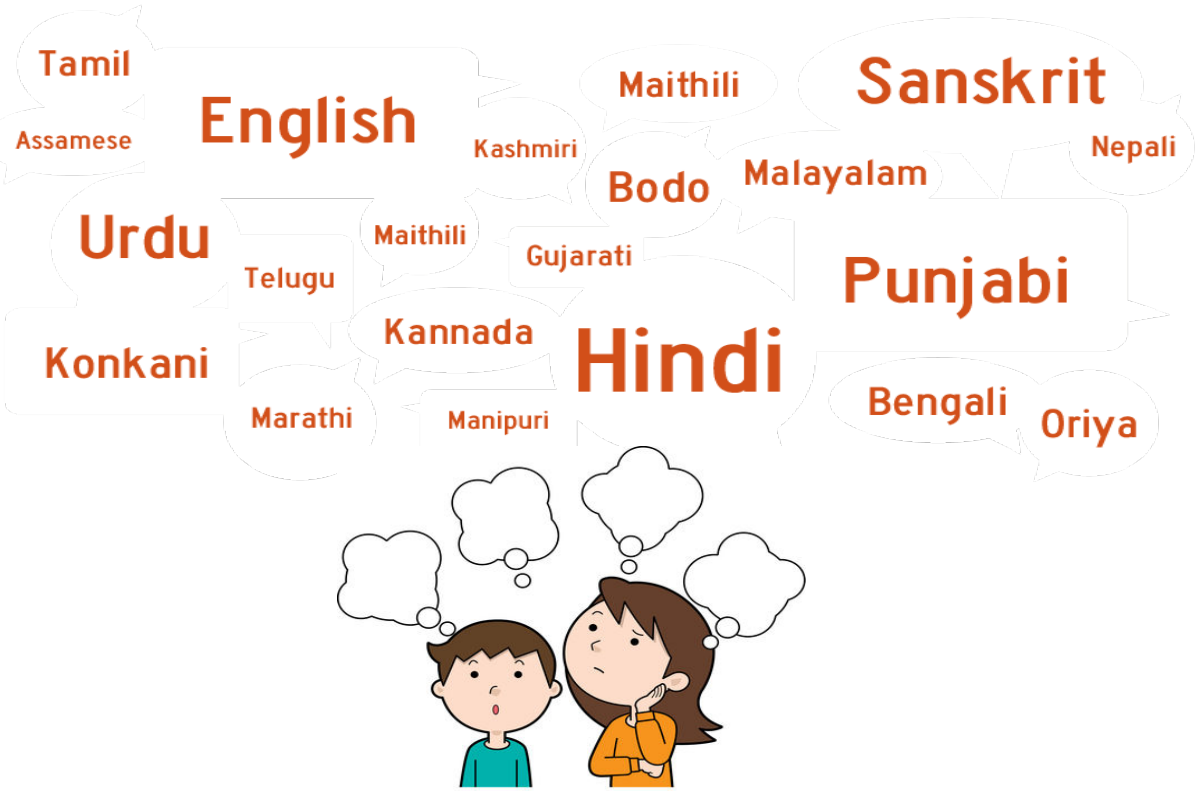
**Collaborators**: : Neha Prabhugaonkar  Kavita Ganeshan, Unnikrishnan Sureshkum

**Supervisor**: Dr. Soudip Roy Chowdhury

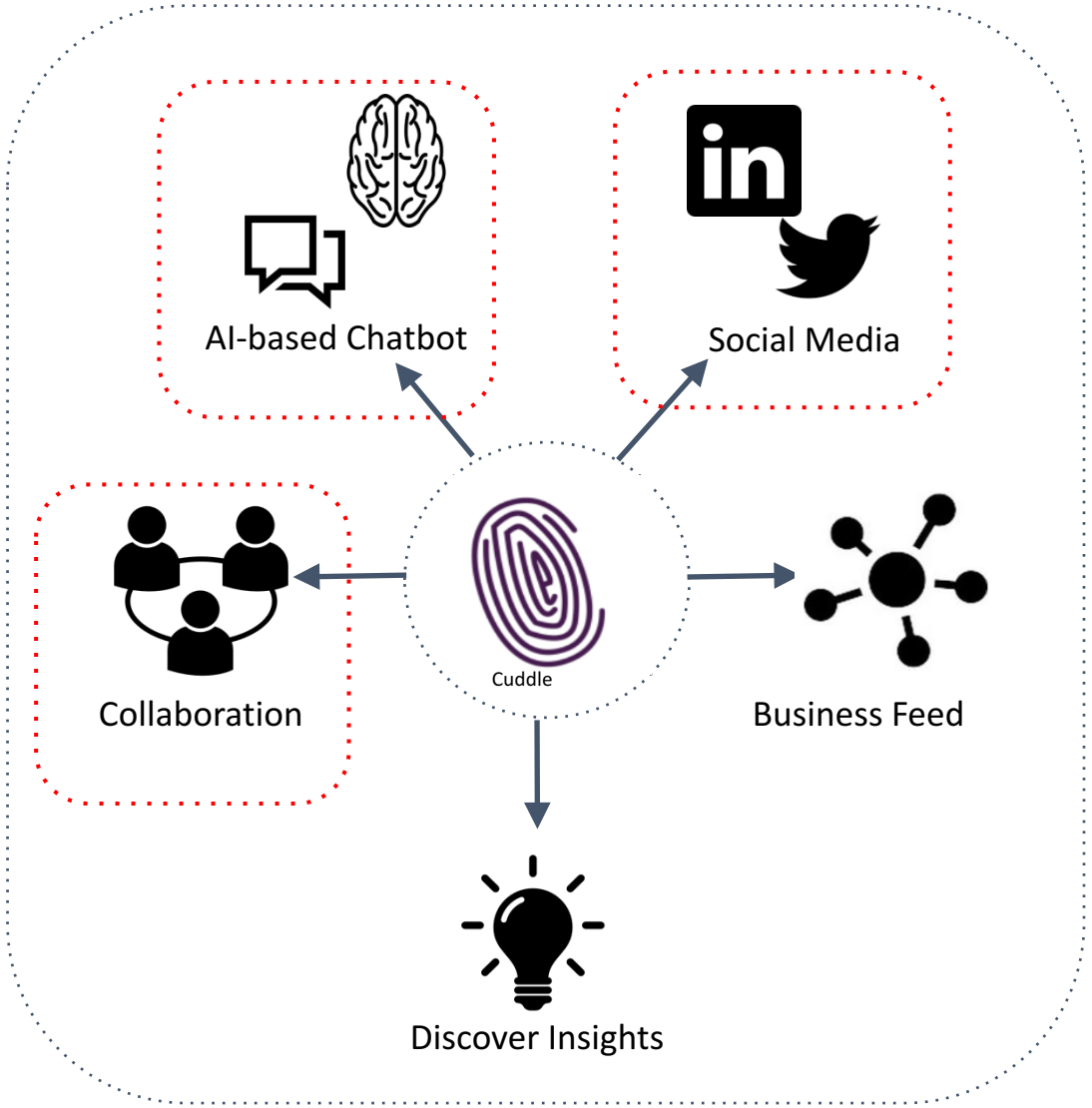
Cuddle Inc.

# Outline

- Motivation
- Code Borrowing in Chat logs
- System Architecture
- Models
- Results
- Future Work
- References



# Motivation


AI-based Chatbot
Social Media
Collaboration
Cuddle
Business Feed
Discover Insights

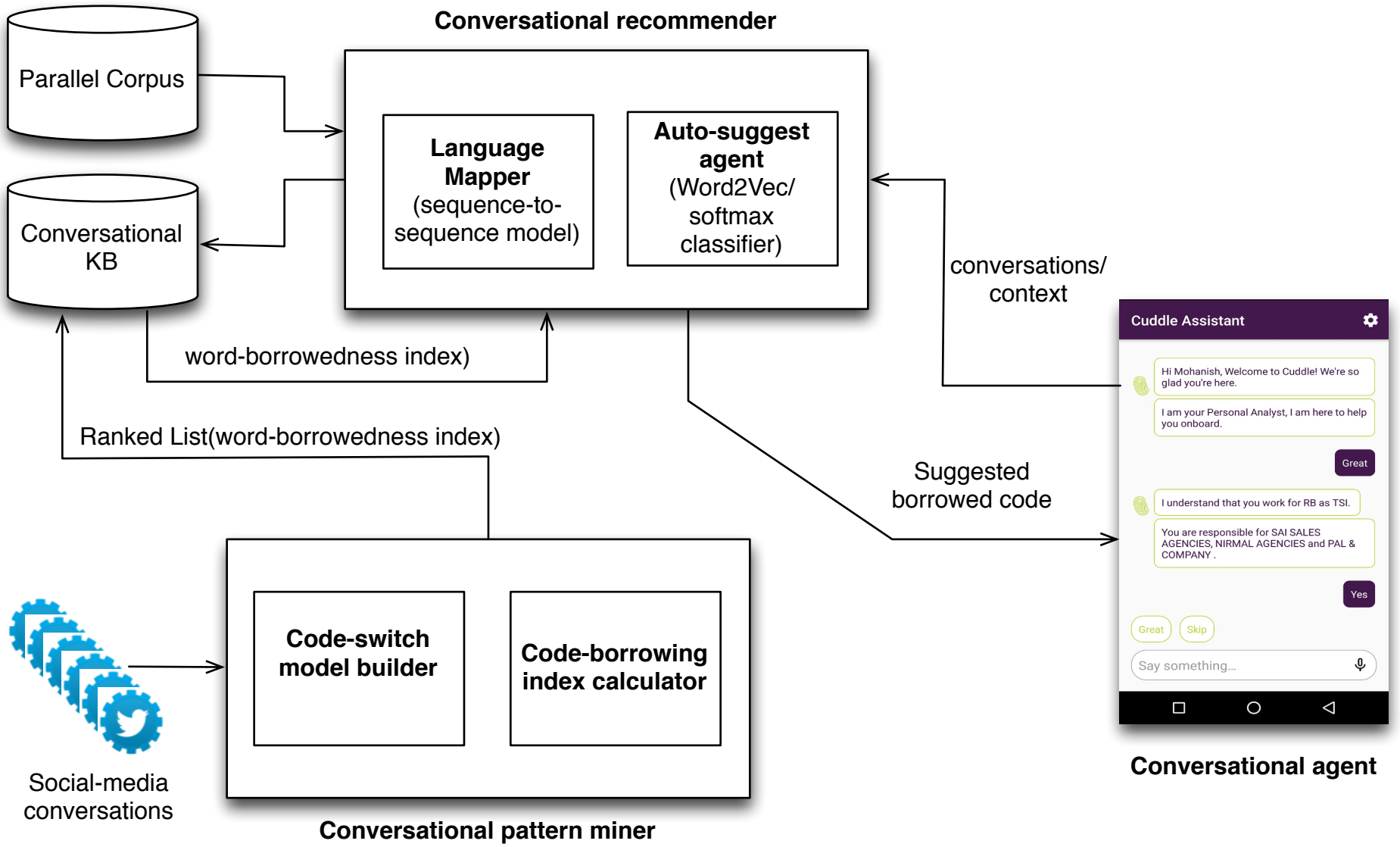
# Code Borrowing in Chat logs


Code mixing

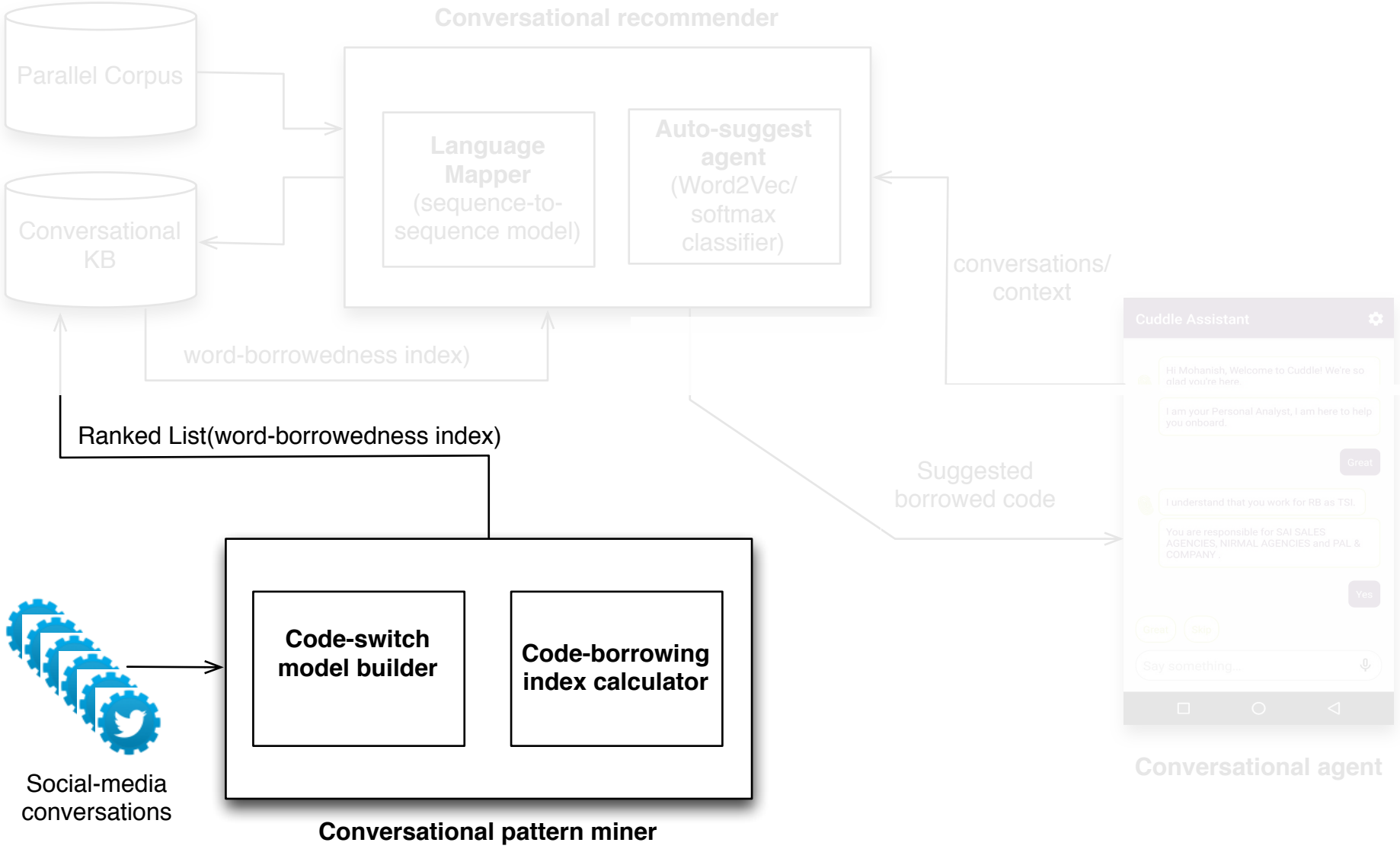

Code borrowing

- Code-mixing leads to Code-borrowing
- Determining the *likelihood* of a code to be borrowed from Code-mixing pattern
  - *borrowedness index*

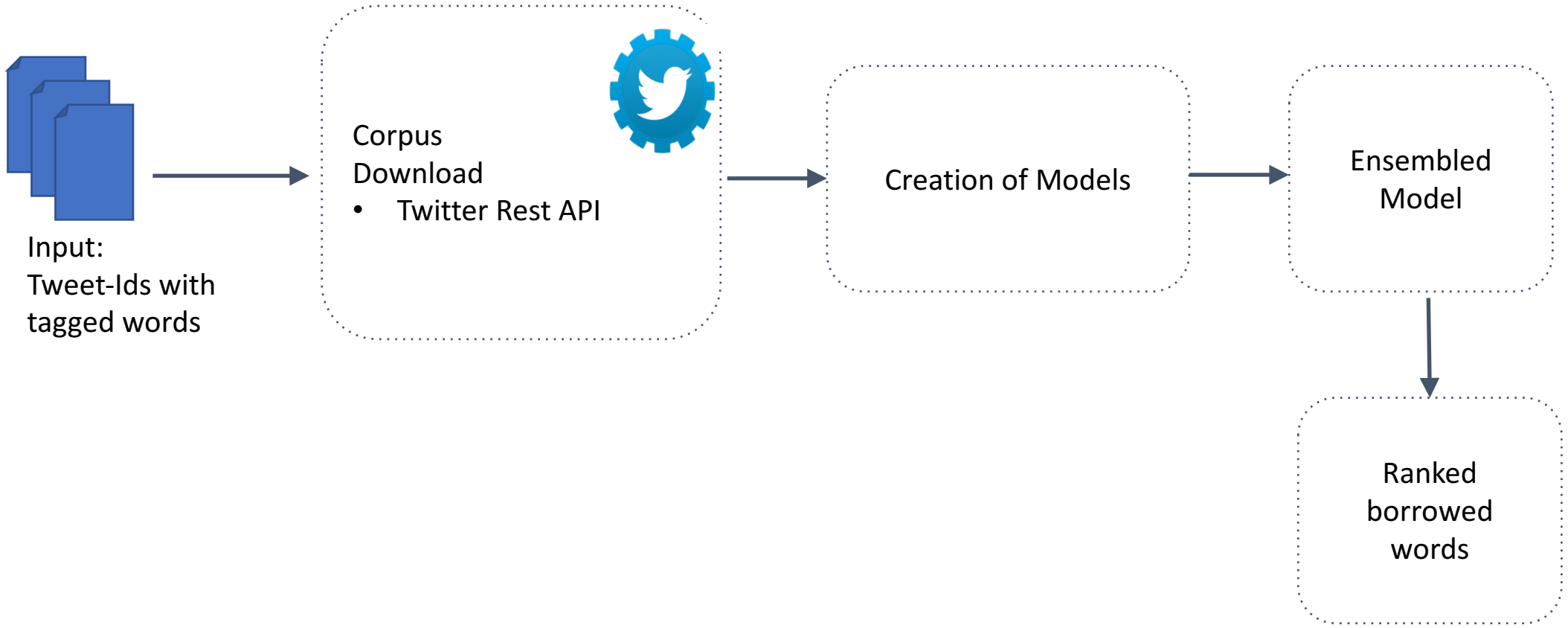
# System Architecture



# In Today's Talk



# Data Flow under pattern miner



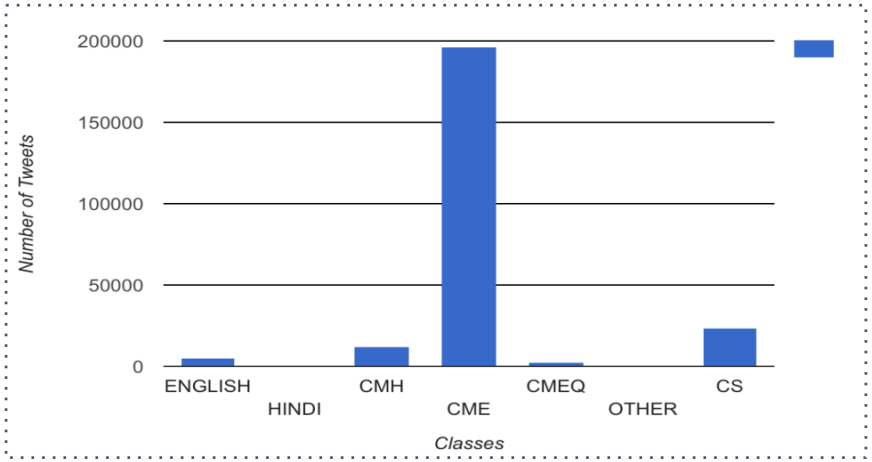
# Chatbot Integration

- Parallel Corpus used for building conversational knowledge [IITB English-Hindi Corpus] (http://www.cfilt.iitb.ac.in/iitb_parallel/)
- Two deep learning-based methods are used
  - Softmax classifier (continous bag of words CBOW) i.e., probability of the next word ("target") given the previous words (“Context”)
  - Sequence-to-sequence model for language mapping
- If the translated target word is ranked high wrt *borrowedness index*
  - Recommend users *translated target word*

# Data Collection

- **Tweet Annotation**
  - The total number of tweets extracted were 240360
  - The distribution of different types of tweets

| Classes | Number of Tweets |
|---|---|
| ENGLISH | 5652 |
| HINDI | 84 |
| CMH | 11950 |
| CME | 196143 |
| CMEQ | 2674 |
| OTHER | 2 |
| CS | 23855 |



# Model Creation

- Model 1: UUR Model
  - Unique User Ratio of a word *w* : *UUR(w) = (U_hi + U_cmh) / U_en*

- Model 2: UTR Model
  - Unique Tweet Ratio of a word *w* : *UTR(w) = (T_hi +T_cmh) / T_en*

# Model Creation

- Model 3: Inverse Model
  - *IM(w) = 0.5 * IUUR(w) + 0.5 * IUTR(w), where*
  - *IUUR(w) = 1 + log(U/UUR)*, where *U = U_hi + U_cmh + U_en*
  - *IUTR(w) = 1 + log(T/UTR)* , where *T = T_hi + T_cmh + T_en*

- Model 4: TF-IDF Model
  - This gives the importance of the word in the corpus

# Model Creation

- Model 5: Weighted Class Average Model (WCAM)
  - *WCAM(w) = [ (PH)*0.25 + (EP)*0.25 + [(SE)+(SH) ]* 0.2 + (CS)*0.1 ] / (PE)*

| | |
|---|---|
| PH | Pure Hindi |
| EP | English Phrase |
| SE | Start English |
| SH | Start Hindi |
| CS | # Code Switch >2 |
| PE | Pure English |

- Model 6: Code-Switched Model
  - *CSR(w) = T_cs/N_en, where N_en = number of times the word is tagged as English*
  - *T_cs = number of CS tweets in which word w is present*

# Results

- Spearman's Rank Correlation

| | | |
|---|---|---|
| **Ground Truth (12 words)** | **UUR + UTR Model** | 0.61 |
| | **Ensemble Model** | 0.63 |

| | | |
|---|---|---|
| **Ground Truth (70 words)** | **UUR + UTR Model** | 0.59 |
| | **Ensemble Model** | 0.62 |

# Results

- Final Ranked List

| word | UUR Model | UTR Model | WCAM Model | Inverse Model | TF-IDF Model | CS Model | Ensemble Model *** |
|---|---|---|---|---|---|---|---|
| **sir** | 7 | 7 | 45 | 117 | 8 | 20 | 1 |
| **main** | 1 | 1 | 3 | 182 | 70 | 1 | 2 |
| **film** | 14 | 13 | 9 | 161 | 23 | 4 | 3 |
| **picture** | 17 | 17 | 19 | 193 | 128 | 16 | 4 |
| **song** | 43 | 55 | 27 | 91 | 39 | 27 | 5 |

***** Runner up at Data Challenge organised by Microsoft research and IIT KGP in COMAD and CODS conference in March, 2017**

# Future Work

- Applying algorithms on various types of data
  - FB, work-related tweets etc.
  - Experiments with European languages
- Applying different algorithms
  - Combining WCAM and TFIDF to get better insights
  - Using RankNet for ranking of borrowed words
- Improving the user-experience of chat bot
  - Measuring the effect of multi-lingual auto-suggest via user engagement

# References


1. Bali et al. (2014): Kalika Bali Jatin, Sharma , Monojit Choudhury, and Yogarshi Vyas. ""“I am borrowing ya mixing?” An Analysis of English-Hindi Code Mixing in Facebook." EMNLP 2014 (2014): 116.
2. Gella et al, (2013): Spandana Gella, Jatin Sharma and Kalika Bali. Query word labeling and Back Transliteration for Indian Languages: Shared task system description In Proceedings of the Fifth Workshop on Forum for Information Retrieval (FIRE 2013). New Delhi, India
3. Gualberto A. Guzman, Jacqueline Serigos, B.E.B.A.J.T. Simple Tools for Exploring Variation in Code-Switching for Linguists. 12–20.
4. Dietterich, T. G. Ensemble methods in machine learning. In International workshop on multiple classifier systems, Springer (2000), 1–15.
5. Data Challenge details: https://ikdd.acm.org/cods2017/data-challenge.html

# Thank You

Questions?