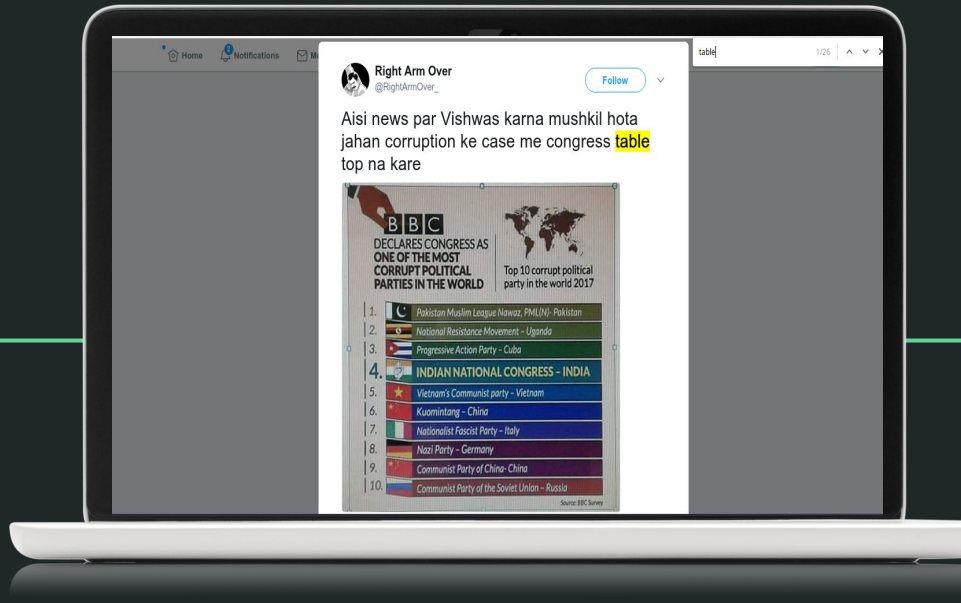
Quantifying sense deviation in Twitter

A study observing online social media tweets
Project 16, Group 6



Problem Statement

- Quantification of sense deviation of a word in social media to conventional English.

- Understand the senses in which a particular word is borrowed.

- If “second” is used in social media, does it mean “दूसरा” or “तुरंत” or “सेकंड” (time) ?

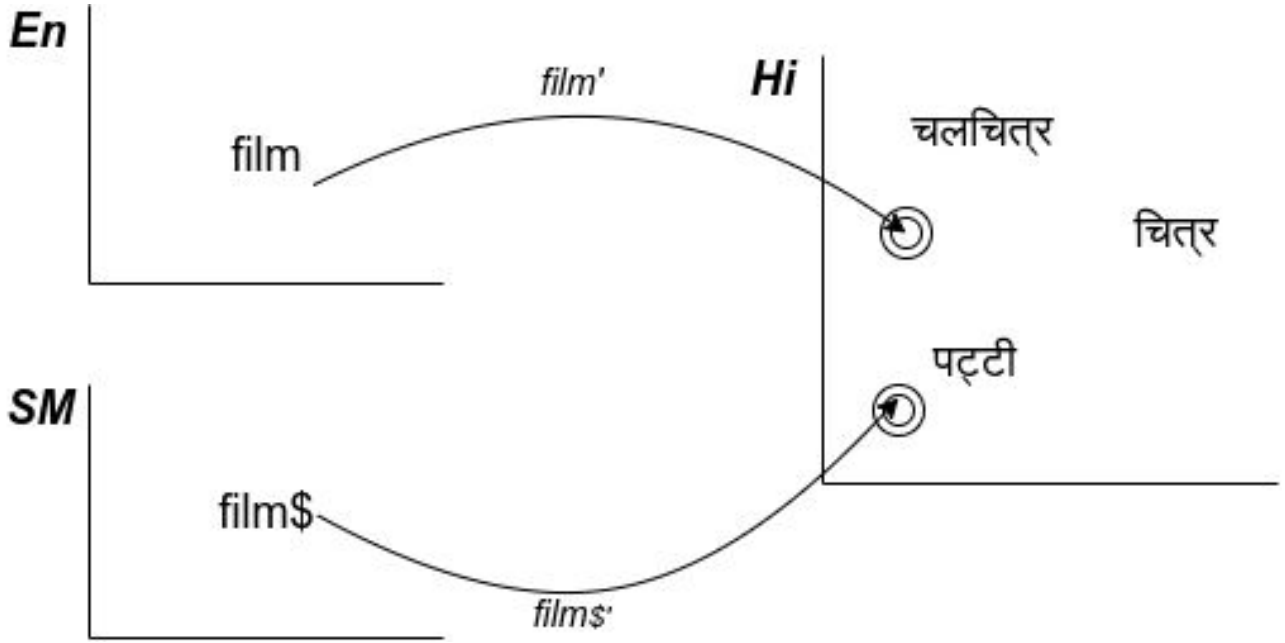
Initial processing, on partial vs complete dataset

Word size	Jacard	spearman-utr	spearman-upr
100 (initial)	0.074	0.252	0.31
100	0.10214	0.803966	0.24725
200	0.10214	0.56010	0.41714

Representing social media dataset

- We obtained possible senses in Hindi for every English word in our dataset.
- For every English word which appears in Hindi context, identify what sense the word is used in.
- To distinguish word in English context and Hindi context, we add a ‘$’ sign to the English word in hindi context.
- “Aaj maine woh star ko dekha” becomes “Aaj maine woh star$ ko dekha”

Comparison of transformed vectors



Visualising social media with Word2Vec

Word2Vec is building word projections (*embeddings*) in a **latent space** of N dimensions, (N being the size of the word vectors obtained).

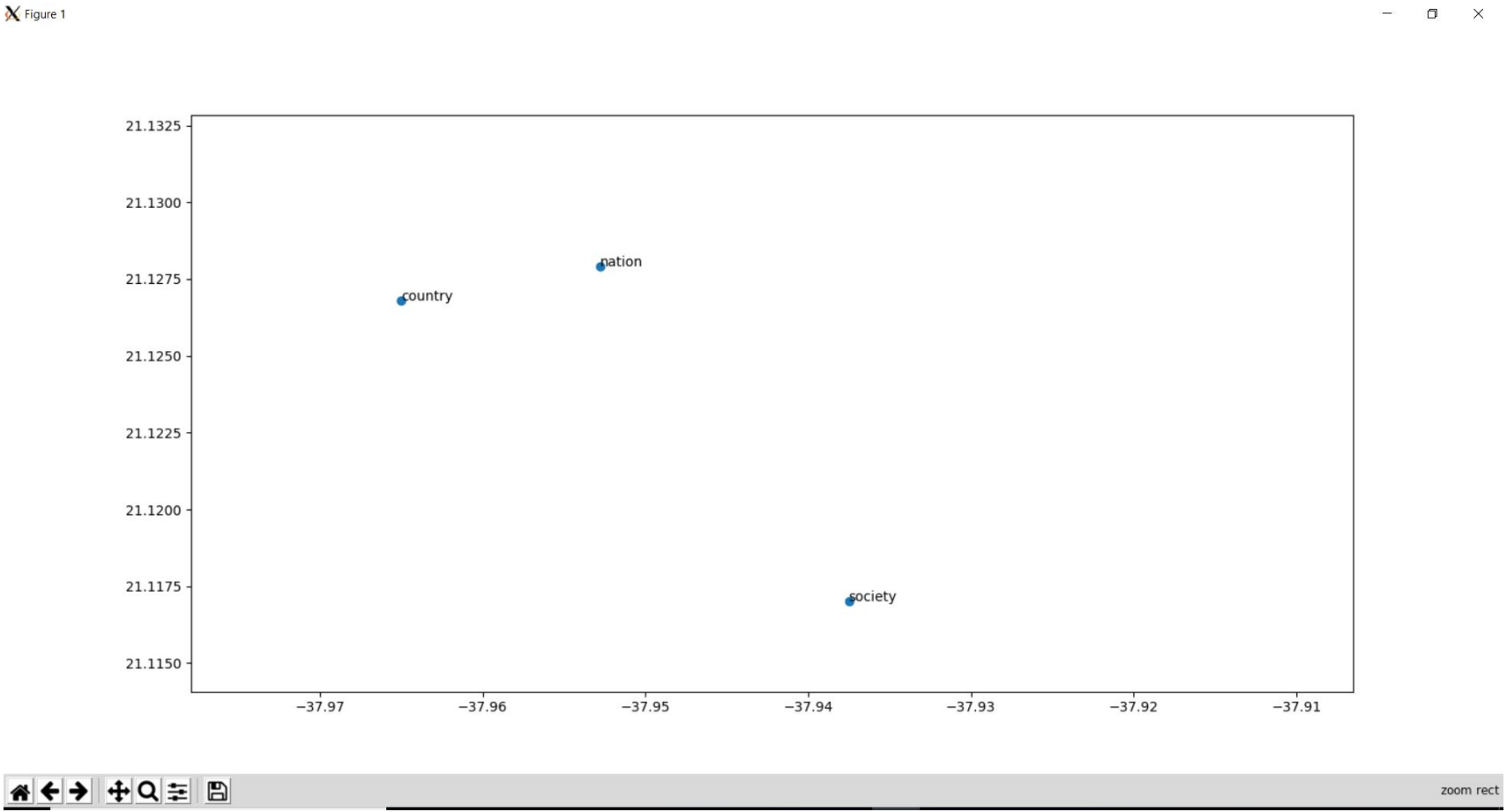
- Obtained coordinates of all words in 300 dimensions.

- Used Dimensionality reduction (TSNE) to visualise vectors in 2 dimensions.

- Stored these points and plotted using Matplotlib

Reference: https://www.tensorflow.org/tutorials/word2vec and https://radimrehurek.com/gensim/models/word2vec.html

Zoomed in view shows clusters being formed



Manual annotation

- A total of 2238 English words in Hindi context were analysed across 21000 tweets.
- Converted English words in Hindi context to word’$’

- Every word was tagged by 2 people

- Inter-annotator agreement
  Cohen’s kappa value was found to be 0.633

Social media dataset

- Analysed 30 lakh tweets which use English and Hindi words.
- Identified 6846 English words
  - 1770 appear both in Hindi and English context.
  - Found 2231 English words in Hindi context
  - Identified 1844 English words in Hindi context with at least a single sense.
- 388 words have multiple meanings, 1456 have single meaning.
- Of the 388 words, 71 words have the transliterated word e.g “सर्विस” as the best sense of “service”
- Of 1456, 154 words have desk: ['डेस्क'] as the only understanding.

Why English word was the best sense?

- Proper nouns, England, indonesia, etc
- Desk, channel
- Technical: internet, site, keyboard, calculation
- Film, shoot, scene, camera, professor
- Some words borrowed from hindi to english as well: Yoga, dacoity,

Transformation of word embeddings

- Vectors from one word2vec model are mapped to another model
- Word, translation pairs from one model to the other used for alignment
- Obtain transformation matrix by minimising the distance between the transformed model and the base model

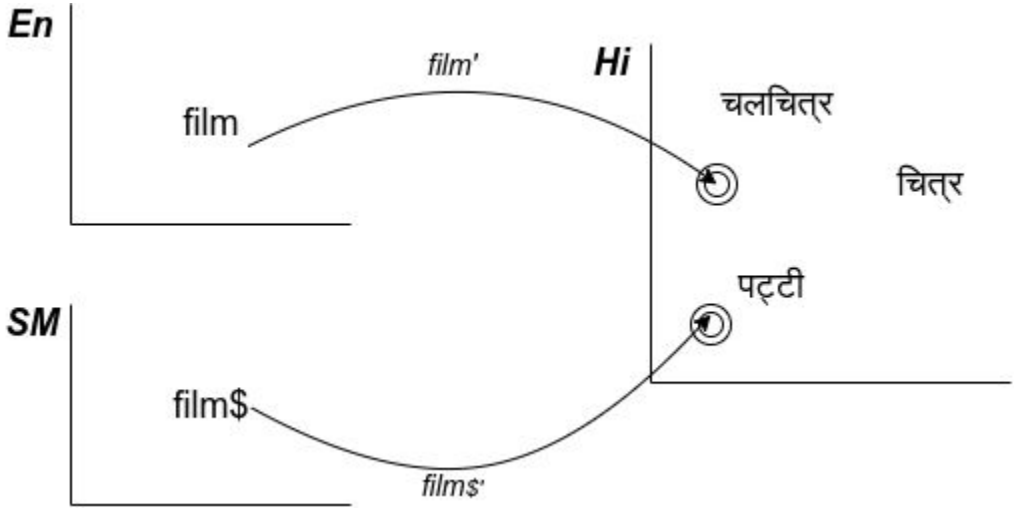
$$\min_{W} \sum_{i=1}^{n} ||Wx_i - z_i||^2$$

- Procrustes alignment method

Reference: Hamilton's (2016): Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change

Comparing word embeddings

- English word2vec model to Hindi word2vec model (say, film)
- Social Media word2vec model to Hindi word2vec model (film$)
- From word annotations, we obtain word, translation pairs (xi, zi)



Results

Choosing Test dataset

- Based on the occurrences of an English word in Hindi and English newspaper corpora, we chose 57 words, the word sense pair of which were excluded from training data.

- Based on the transformed vectors, we consider the following 2 results:
  - Comparing cosine similarity of a sense in English and social media
  - Finding cosine similarity of word in conventional English and social media

Result 1: Comparing similarity among senses

Cosine similarity difference = cos(Eng, Hin) - cos(Soc, Hin)

English word	Social Media	Hindi sense	cosine difference	Hindi sense	cosine difference
Star	Star$	तारा	0.09824002109	सुपरस्टार	-0.180909364
Scene	Scene$	स्थल	0.2753678874	शॉट	-0.3110012461
well	well$	अच्छा	-0.1046849495	काफी	0.02596171899

Result 2: Similarity between transformed vectors

Once transformed, how similar are those 2 words?

Film and film$

If cosine similarity is high, it means the word is used in similar senses in both sm and en. If low, it means film$ used in different contexts than film.

Result 2: Similarity between transformed vectors

- We found for words with singular sense, the value for cosine similarity was positive and high.

English	SM	Cosine similarity
god	god$	0.2421904787
blue	blue$	0.127730732
president	president$	0.1222979614
woman	woman$	-0.03398525207
job	job$	-0.1183152846
play	play$	-0.1574467486

Conclusion

- Implemented transformation matrix using Hamilton (2016) (Procrustes method) from scratch.

- We found certain words are closer to a sense than when used in English.

- We need more data to fully understand the trends and sense distribution for English words in social media.

Thank you !

Group 6

Kaustubh Hiware	14CS30011
G. Prithvi Raj Reddy	14CS10016
Kiran Sing Sastry G.	14CS10018
T. Karthik	14CS10049
Surya M.	14CS30017

Result 2: Comparing nearest neighbors

For each word in social media, get 10 nearest neighbors for each word

For transformed vector

Out of 54 cases, 54 English words had at least one neighbour relevant to the word.