Project Topic:

Contextual Embeddings (ELMo) for Indian Languages

&

A Novel Semi-Intrinsic Approach to Evaluate Word Embeddings for Robustness

Gaurang Patil – 2023701002

Ishank Kapania – 2023701012

Anshul Sharma – 2023701011

Problem Statement & Introduction

- Problems with non-contextual embeddings like Word2Vec
- Contextual embeddings: ELMo (for Hindi)
- Evaluation of contextual embeddings using extrinsic tasks
- Creation of new dataset and proposal of a semi-intrinsic approach to evaluate the quality and robustness of contextual word embeddings

ELMo

$$
\begin{aligned}
R_k &= \{\mathbf{x}_k^{LM}, \overrightarrow{\mathbf{h}}_{k,j}^{LM}, \overleftarrow{\mathbf{h}}_{k,j}^{LM} \mid j = 1, \ldots, L\} \\
&= \{\mathbf{h}_{k,j}^{LM} \mid j = 0, \ldots, L\},
\end{aligned}
$$

$$
\mathbf{ELMo}_k^{task} = E(R_k; \Theta^{task}) = \gamma^{task} \sum_{j=0}^{L} s_j^{task} \mathbf{h}_{k,j}^{LM}.
$$

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Issues with word similarity

- Inability to handle polysemy
  - e.g. जल - पानी / जलना ?
  - Cannot handle contextual embeddings

- Unfairly penalize embeddings trained for specific tasks
  - Word representations trained for specific tasks e.g. POS tagging (and not co-occurrence prediction) capture *task specific* word similarity (and not semantic similarity

  - For POS tagging task – nouns 'cat' and 'man' might be considered similar by word model, but word similarity would unfairly penalize them

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Issues with word similarity

- Low correlation with extrinsic evaluation
  - No strong correlation between performance on word similarity and extrinsic tasks like classification
  - Authors call for alternative approaches to evaluation
- Frequency effects in cosine similarity
  - Formation of hubs in vector spaces: vectors are close to a large number of other vectors, resulting in higher cosine similarity between words
  - Words with similar frequency are closer in embedding space – unfairly assigns higher similarity
- More issues – but our proposed approach does not provide a solution for the same

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 30–35, Berlin, Germany. Association for Computational Linguistics.

Issues with intrinsic evaluation for low resource languages

- Datasets
  - Mostly comprise translations of English word similarity datasets
  - Intrinsic evaluation often cannot be applied in contextual settings. E.g. word similarity and analogy.

- However, several Indic languages have extrinsic evaluation datasets
  - Hate speech, sentiment classification, NER, POS tagging are commonly available

Word Level Adversarial Attacks

- Authors attribute vulnerability to word level adversarial attacks to the fact that similar words get mapped to dissimilar representations

- Fast Triplet Metric Learning (FTML)
  - Pull words closer to their positive samples (i.e. synonyms)
  - Push words away from negative samples (i.e. non-synonyms)

- Creates robust embeddings

- Key idea : "*A robust classifier should be able to extract similar representations when fed with similar input samples*"

Yang, Y., Wang, X. & He, K.. (2022). Robust textual embedding against word-level adversarial attacks. *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, in *Proceedings of Machine Learning Research* 180:2214-2224 Available from https://proceedings.mlr.press/v180/yang22c.html.

Issues with extrinsic evaluation

- Extrinsic evaluation can be helpful for high-level comparison of word embeddings
- May not provide detailed and interpretable information on various nuances of word embedding quality
- Deeper analysis of word embeddings is not possible
  1. Deeper analysis can provide useful insights to linguists in morphologically and lexically rich languages
  2. Actionable insights : Can be further used to understand how to improve the word model itself

Need benchmarks other than human judgment of similarity

"Moreover, it is important to propose benchmarks other than human judgments of similarity, i.e., benchmarks for in-vivo evaluation strategy based on the analysis of the performance of applications which rely on semantic measures"

Hadj Taieb, M.A., Zesch, T. & Ben Aouicha, M. A survey of semantic relatedness evaluation datasets and procedures. A*rtif Intell Rev* **53**, 4407–4448 (2020). https://doi.org/10.1007/s10462-019-09796-3

Can we still measure "*word similarity*" or robustness using a semi- intrinsic approach and yet overcome the issues discussed to a good extent?

Intuition / Core idea

"Replacing a word with its synonym should not cause a model to perform worse on a downstream task"

"The cost incurred by substituting a word with its synonym indirectly measures the quality of word embeddings"

Note: The substitution need not always be a synonym, it depends upon the downstream task

Datasets

- IndicGLUE
- IndicNLP corpora
- IndicXNLI
- HiNER
- Self-created dataset using Product Reviews dataset from IndicGLUE as the base

Summary of Implementation

- **Training LMs & Extrinsic evaluation:**
  - Training of the forward and backward LMs
    - Train on a generic corpus using 1 lakh sentences on next word prediction and previous word prediction.
  - Creation of task specific ELMo embeddings for each of the tasks
  - Train biLSTM for the downstream tasks
  - Evaluate the performance of the model on the downstream tasks
  - For the MCQ's based dataset we concatenated the Question and the respective Questions to one fixed tensor that was fed to a downstream LSTM as suggested in the IndicGLUE paper. While then the classification was performed on 4 classes meaning 4 options for the MCQs

- **Proposed metric:**
  - Development of a new dataset for semi-intrinsic evaluation of word embeddings
  - Evaluate the word embeddings using the proposed metric
  - Obtain embeddings with frozen lambdas and evaluate their quality using the proposed metric.

Summary of Implementation

- **Computing the metric:**
  - A semi-intrinsic approach to evaluate robustness of word embeddings
  - Create a set of 100 original sentences
  - For each sentence, substitute an “important” word with an appropriate synonym. Make grammatical changes if necessary and obtain the new sentence.
  - Let the classifier obtain y_hat1 and y_hat2 which are the predicted probabilities for the 2 sentences for the target class.
  - Compute | y_hat1 – y_hat2 | for each pair and store it in a list.
  - Median of the list gives the final value of the metric.

Example

- Original sentence: तुम बहुत **घमंडी** हो
- After careful substitution : तुम बहुत **अहंकारी** हो
- Say y_hat = 0.8 and y_hat_syn = 0.77 for target class
- Substitution with synonym has costed us
  - A penalty of: 0.8 - 0.77 = 0.03
- Notice:
  - Human assigned scores for word similarity are not required
  - Non-reliance on cosine similarity

High Level Overview of Results

Task	Our Test Accuracy	Highest Test Accuracy mentioned in IndicGLUE paper
Product Review Classification	73%	78.97 % (XLM-R)
Movie Review Classification	54 %	61.61 % (XLM-R)
BBC news classification	67 %	75.52 % (XLM-R)
MIDAS discourse mode classification	69 %	79.94 % (XLM-R)
Cloze-style Multiple-choice QA	33%	41.55 % (IndicBERT base)
Wikipedia Section Title Prediction	24 %	80.12 % (mBERT)

High Level Overview of Results

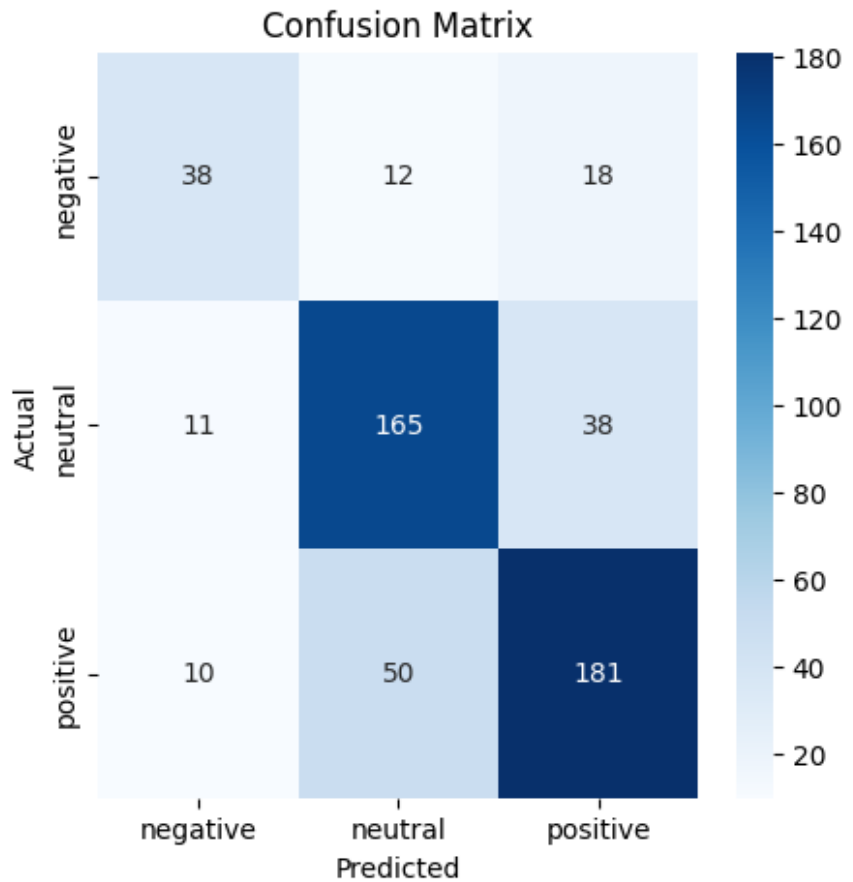
Dataset	**Task**	**Our Test metric**	**Highest Test Metric mentioned in paper**
HiNer (Published in ACL 2022)	**Named entity recognition**	70% (Macro avg f1-score)	86.98 ± 0.22% (Macro avg f1-score XLM-R large)
IndicXnli (Published in EMNLP 2022)	**Natural language inference**	55.1% (Test accuracy)	78% (Test accuracy XLM-R)

Results: Proposed Metric

Setting	**Value of the proposed metric (Mean)**	**Value of the proposed metric (Median)**
Trainable Lambdas	0.0820	0.0178
Frozen Lambdas	0.0907	0.0267
Original Embeddings	0.0621	0.0229

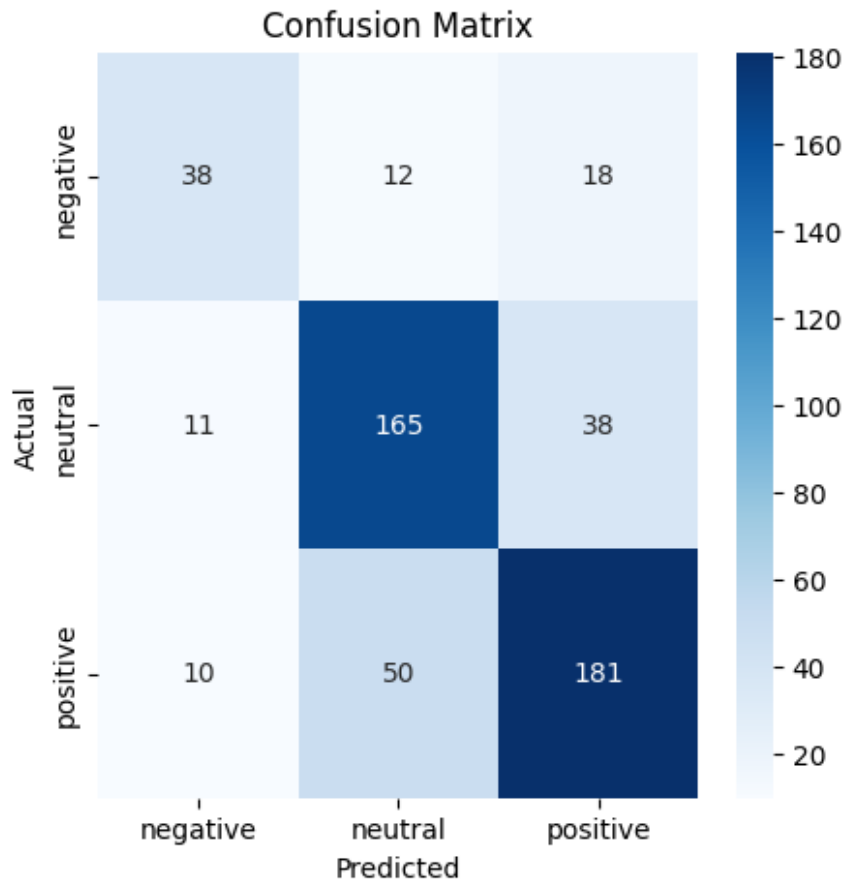
Results- Product Review Sentiment Classification

	precision	recall	f1-score	support
negative	0.64	0.56	0.60	68
neutral	0.73	0.77	0.75	214
positive	0.76	0.75	0.76	241
accuracy			0.73	523
macro avg	0.71	0.69	0.70	523
weighted avg	0.73	0.73	0.73	523



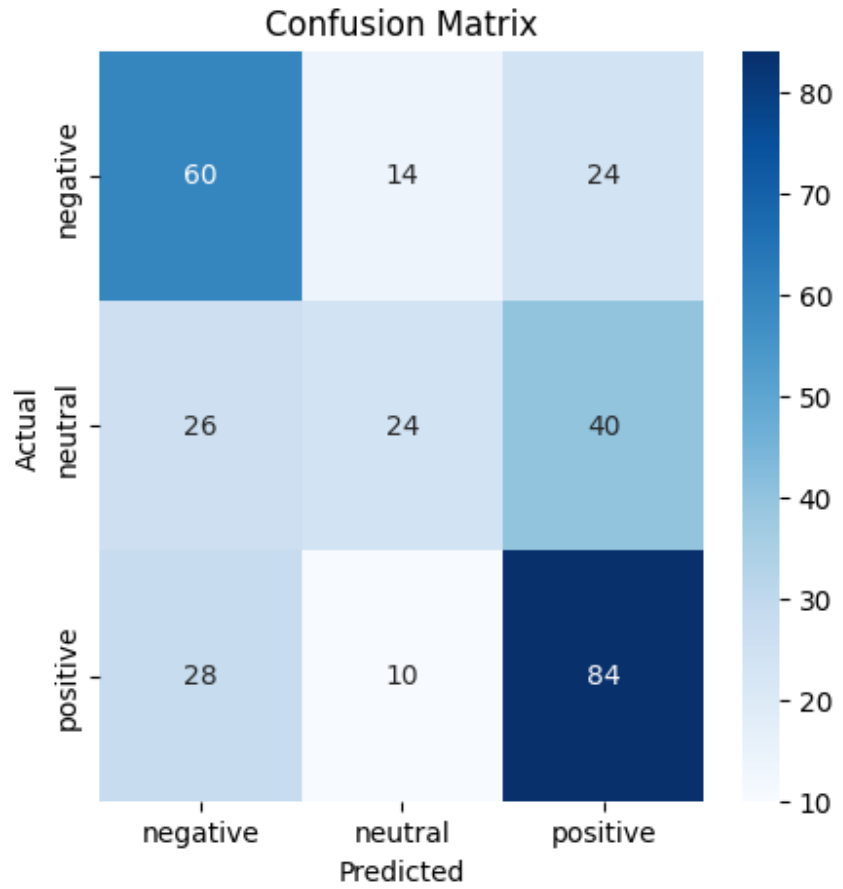
Results- Product Review Sentiment Classification

	precision	recall	f1-score	support
negative	0.64	0.56	0.60	68
neutral	0.73	0.77	0.75	214
positive	0.76	0.75	0.76	241
accuracy			0.73	523
macro avg	0.71	0.69	0.70	523
weighted avg	0.73	0.73	0.73	523



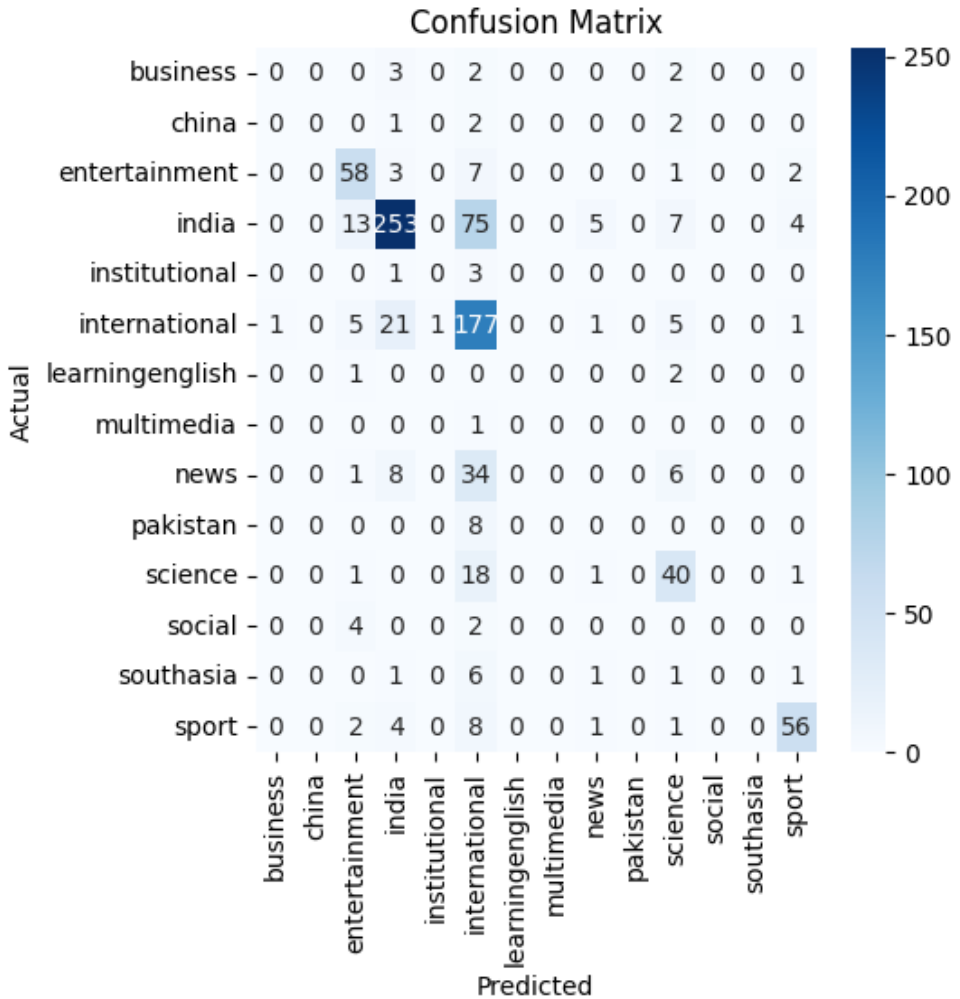
Results- Movie Review Sentiment Classification

	precision	recall	f1-score	support
negative	0.53	0.61	0.57	98
neutral	0.50	0.27	0.35	90
positive	0.57	0.69	0.62	122
accuracy			0.54	310
macro avg	0.53	0.52	0.51	310
weighted avg	0.53	0.54	0.52	310



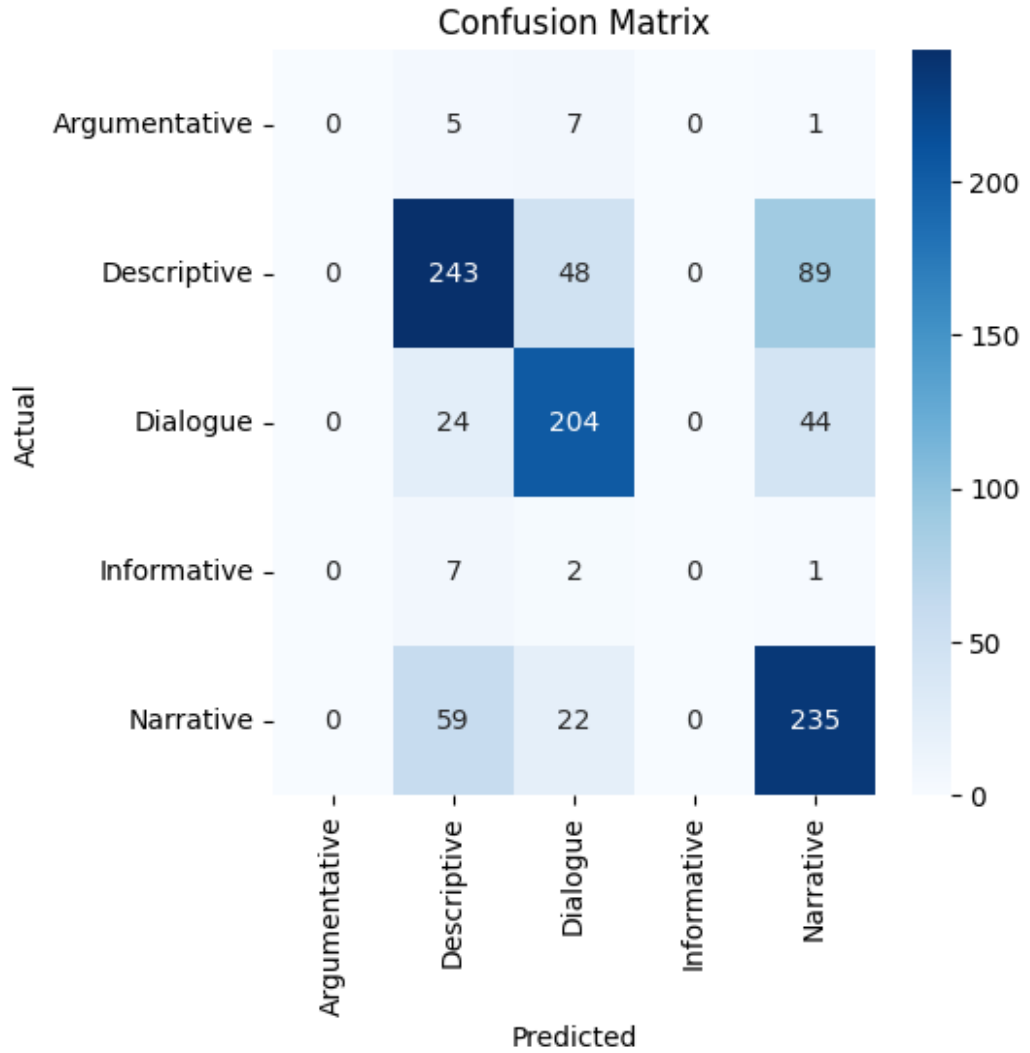
Results- BBC News Category Classification

	precision	recall	f1-score	support
business	0.00	0.00	0.00	7
china	0.00	0.00	0.00	5
entertainment	0.68	0.82	0.74	71
india	0.86	0.71	0.78	357
institutional	0.00	0.00	0.00	4
international	0.52	0.83	0.64	212
learningenglish	0.00	0.00	0.00	3
multimedia	0.00	0.00	0.00	1
news	0.00	0.00	0.00	49
pakistan	0.00	0.00	0.00	8
science	0.60	0.66	0.62	61
social	0.00	0.00	0.00	6
southasia	0.00	0.00	0.00	10
sport	0.86	0.78	0.82	72
accuracy			0.67	866
macro avg	0.25	0.27	0.26	866
weighted avg	0.65	0.67	0.65	866



Results- MIDAS Discourse Mode Classification

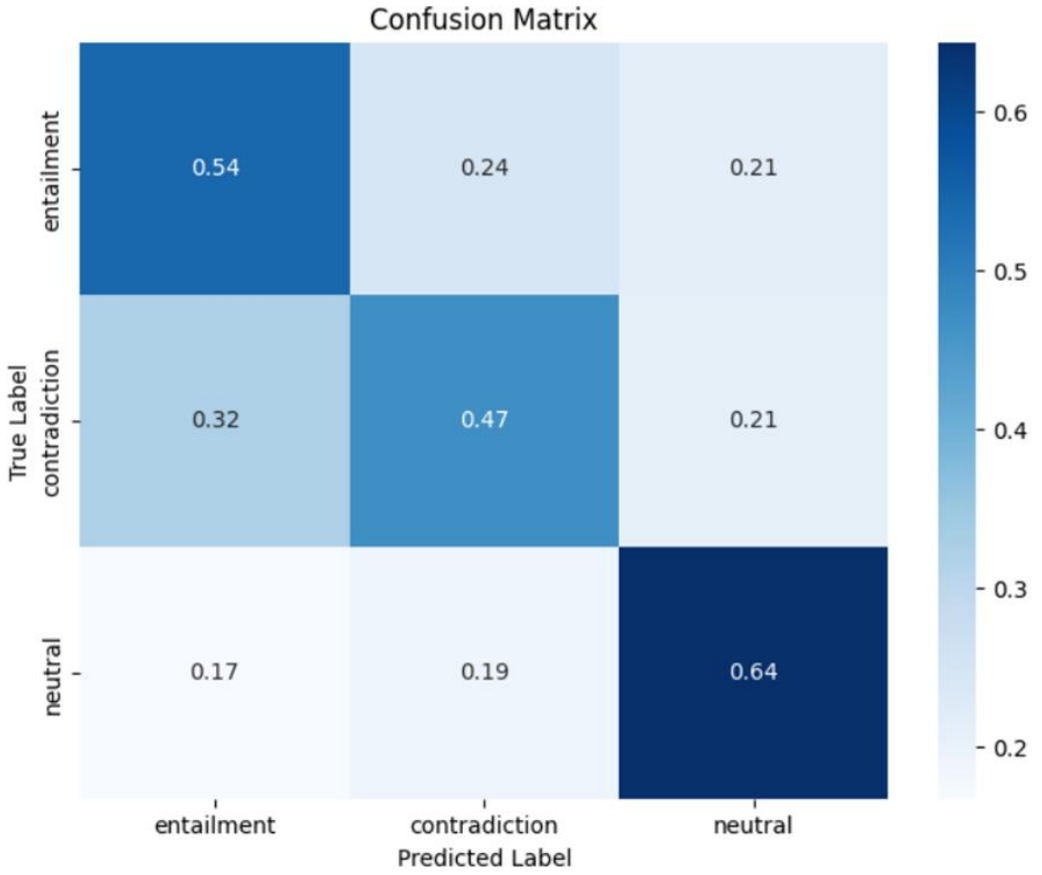
	precision	recall	f1-score	support
Argumentative	0.00	0.00	0.00	13
Descriptive	0.72	0.64	0.68	380
Dialogue	0.72	0.75	0.74	272
Informative	0.00	0.00	0.00	10
Narrative	0.64	0.74	0.69	316
accuracy			0.69	991
macro avg	0.41	0.43	0.42	991
weighted avg	0.68	0.69	0.68	991



Results- NLI

Classification Report:

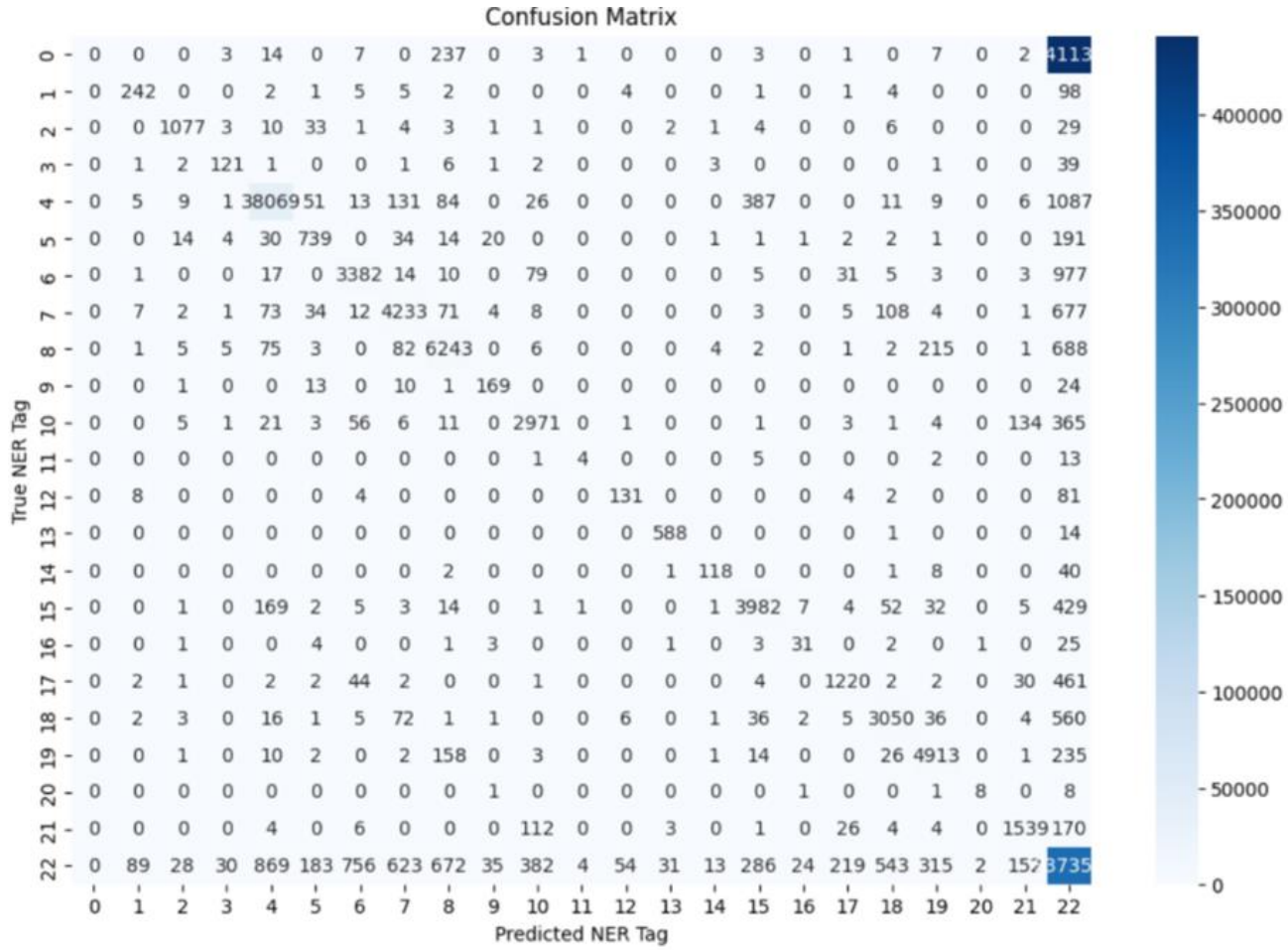
	precision	recall	f1-score	support
entailment	0.53	0.54	0.53	1670
contradiction	0.52	0.47	0.49	1670
neutral	0.60	0.64	0.62	1670
accuracy			0.55	5010
macro avg	0.55	0.55	0.55	5010
weighted avg	0.55	0.55	0.55	5010



Results- NER

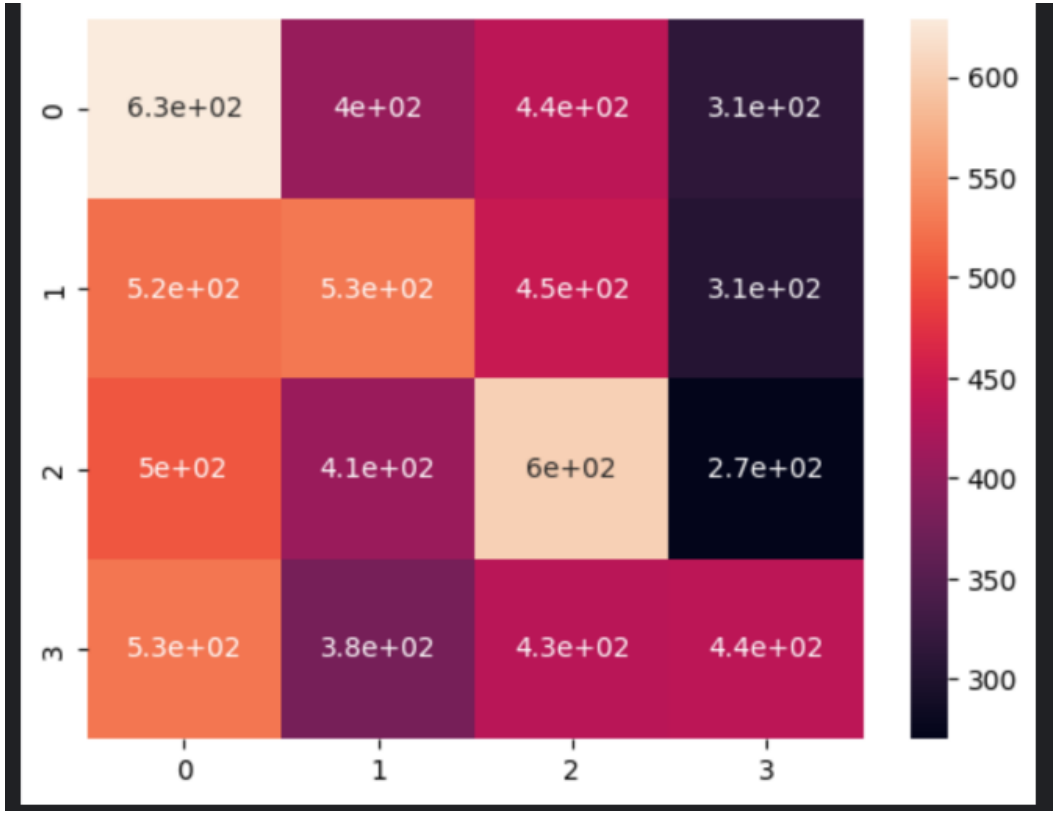
Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	441412
1	0.68	0.66	0.67	365
2	0.94	0.92	0.93	1175
3	0.72	0.68	0.70	178
4	0.97	0.95	0.96	39889
5	0.69	0.70	0.70	1054
6	0.79	0.75	0.77	4527
7	0.81	0.81	0.81	5243
8	0.83	0.85	0.84	7333
9	0.72	0.78	0.75	218
10	0.83	0.83	0.83	3583
11	0.40	0.16	0.23	25
12	0.67	0.57	0.62	230
13	0.94	0.98	0.96	603
14	0.83	0.69	0.75	170
15	0.84	0.85	0.84	4708
16	0.47	0.43	0.45	72
17	0.80	0.69	0.74	1773
18	0.80	0.80	0.80	3801
19	0.88	0.92	0.90	5366
20	0.73	0.42	0.53	19
21	0.82	0.82	0.82	1869
...				
accuracy			0.47	866280
macro avg	0.72	0.71	0.70	866280
weighted avg	0.25	0.47	0.32	866280



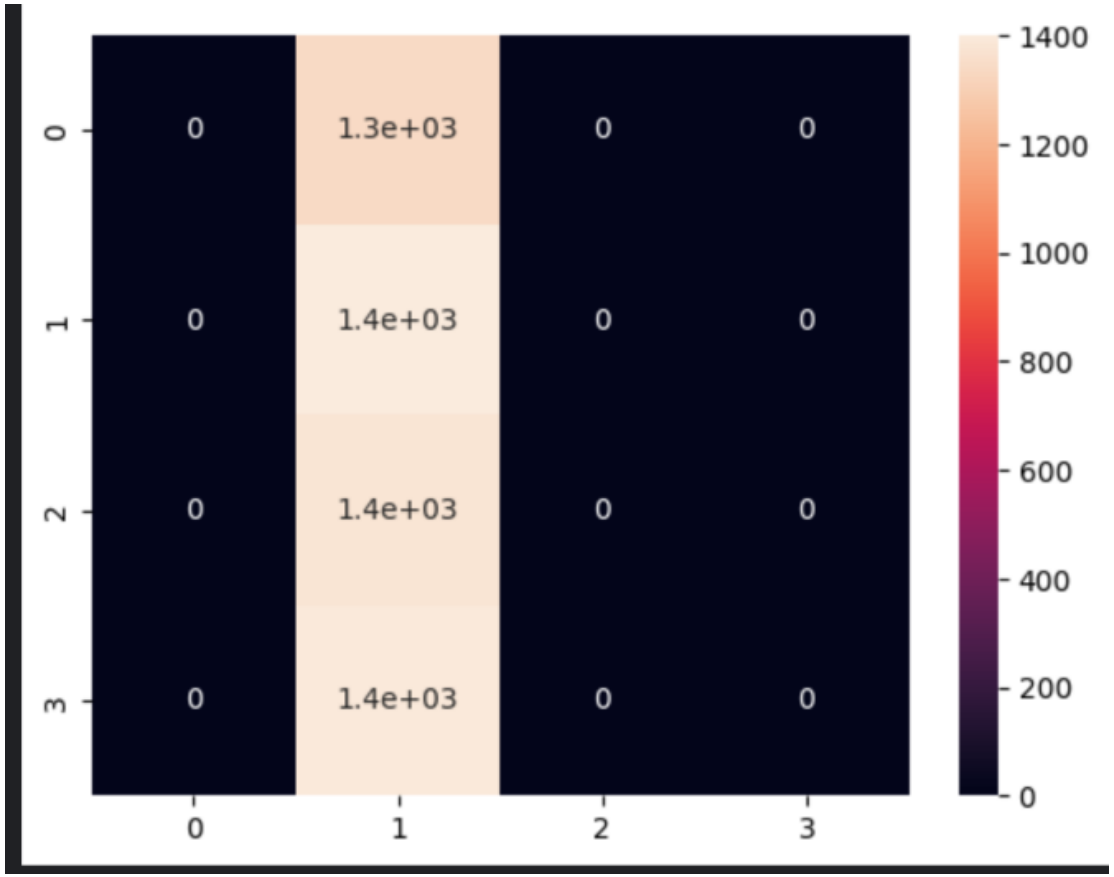
Results : Cloze Style Multiple Choice QnA

	precision	recall	f1-score	support
0	0.30	0.45	0.36	1761
1	0.38	0.17	0.24	1795
2	0.33	0.37	0.35	1826
3	0.34	0.33	0.33	1758
accuracy			0.33	7140
macro avg	0.34	0.33	0.32	7140
weighted avg	0.34	0.33	0.32	7140



Results: Wikipedia Section Title Prediction

	precision	recall	f1-score	support
0	0.24	1.00	0.39	1338
1	0.00	0.00	0.00	1402
2	0.00	0.00	0.00	1376
3	0.00	0.00	0.00	1393
accuracy			0.24	5509
macro avg	0.06	0.25	0.10	5509
weighted avg	0.06	0.24	0.09	5509



Advantages of the Proposed Metric

Ability to handle polysemy:

- तुम उस ***पद*** के अधिकारी नहीं हो
  - Contextual embeddings will be able to figure out what context needs to be used for पद (as opposed to word similarity where पद would have been ambiguous)

- तुम उस ***ओहदे*** के अधिकारी नहीं हो
  - Substitution is done carefully (manually), so no issues here
  - IndoWordnet is to be used to assist the human with synonyms

Advantages of the Proposed Metric

No unfair penalty for task-specific word embeddings which capture task specific word similarity:

e.g. for POS tagging task- बिल्ली and बच्ची are 'similar' because both are nouns

- बिल्ली पानी पी रही है
- बच्ची पानी पी रही है
  - Note that since the task is POS tagging, the substitution does not occur with synonym but with a word which has the same tag
- बिल्ली and बच्ची both are nouns and since we are evaluating it on a downstream task (POS tagging), no unfair penalty is imposed

Advantages of the Proposed Metric

Ability to check robustness against word level adversarial attacks:

- Useful in social media context
  e.g. hate speech classification

Disadvantages of the Proposed Metric

- There are some corner cases where the metric would fail to detect bad embeddings. E.g. when all word vectors are initialized to the same or very similar values.
  - Solution: Apart from the proposed metric, take into consideration the model's performance on the downstream task using those 100 sentences. A bad word model would perform poorly, even if the proposed metric indicates presence of robust embeddings.
- Computational cost is higher
- Requires creation of an additional dataset which needs human resources.
- Dependency on the ability of the trained model.

Other References

- [1] Hadj Taieb, M.A., Zesch, T. & Ben Aouicha, M. A survey of semantic relatedness evaluation datasets and procedures. *Artif Intell Rev* **53**, 4407–4448 (2020). https://doi.org/10.1007/s10462-019-09796-3
- [2] Wang B, Wang A, Chen F, Wang Y, Kuo C-CJ. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing*. 2019;8:e19. doi:10.1017/ATSIP.2019.12
- [3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- [4] Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- [5] Yang, Y., Wang, X. & He, K.. (2022). Robust textual embedding against word-level adversarial attacks. Proceedings of the Thirty-Eighth Conference on Uncertainty in
- Artificial Intelligence, in Proceedings of Machine Learning Research 180:2214-2224 Available from https://proceedings.mlr.press/v180/yang22c.html.

Thank you for your time!