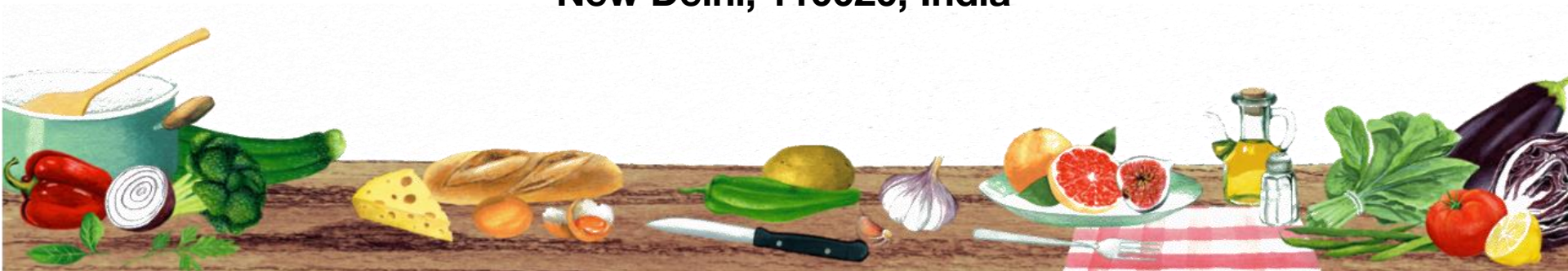
INDRAPRASTHA INSTITUTE of INFORMATION TECHNOLOGY DELHI

LREC-COLING 2024

Deep Learning Based Named Entity Recognition Models for Recipes

Mansi Goel

**Complex System Lab, Department of Computational Biology,
Indraprastha Institute of Information Technology Delhi,
New Delhi, 110020, India**



Named Entity Recognition

- **Food** touches our lives through various endeavors, including flavor, nourishment, health, and sustainability.
- Recipes are cultural capsules transmitted across generations via **unstructured text**.
- Automated protocols for **recognizing named entities**, the building blocks of recipe text, are of immense value for various applications ranging from information extraction to novel recipe generation.
- Named entity recognition is a technique for **extracting information** from unstructured or semi-structured data with known labels.
- NER plays an essential role in various **natural language processing applications** such as text understanding, information retrieval, automatic text summarization, question answering, machine translation, and knowledge base construction, etc.

Literature Review

1. **Popovski et al. 2019** proposed a rule-based NER model **FoodIE**, where the rules incorporate computational linguistics information. FoodIE achieved promising results on independent benchmark datasets and has been used to create the FoodBase corpus, the first NER corpus in the food domain. The **limitation** of the FoodIE method is its dependency on external resources, which have become inaccessible after its publication, rendering the method unusable.
2. **Cenikj et al. 2020** proposed a data-driven method **BUTTER** for named-entity extraction, trained on the FoodBase corpus based on Bidirectional Long Short-Term Memory and conditional random field methods.

Literature Review

3. **Diwan et al. 2020** used the RecipeDB dataset to identify the named entities in ingredient phrases and cooking instructions. They reported an F1 score of 0.95 (ingredient), 0.88 (processes), and 0.90 (utensils).
4. **Radu et al. 2022** implemented NER on cooking instructions from multilingual recipes (French, German, and English). They implemented a Conditional Random Field layer on top of Bidirectional Long-Short Term Memory models, achieving F1 scores over 96% in mono and multi-lingual contexts for all classes.
5. **Cenikj et al. 2022** implemented a BERT-based method, SciFoodNER for recognizing named entities in scientific texts and achieved an F1 score of 0.90.

Motivation

- **Computational Gastronomy** represents the study of food, flavors, nutrition, health, and sustainability from the computing perspectives **[Goel et al. 2022]**.
- This new data science changes the outlook on food and cooking, traditionally considered artistic endeavors.
- Building **NER models for recipe texts** is an exciting proposition, given its applications spanning multiple domains, including disease prediction, cost estimation, flavor profiling, and comprehensive nutritional analysis of recipes.

Contributions

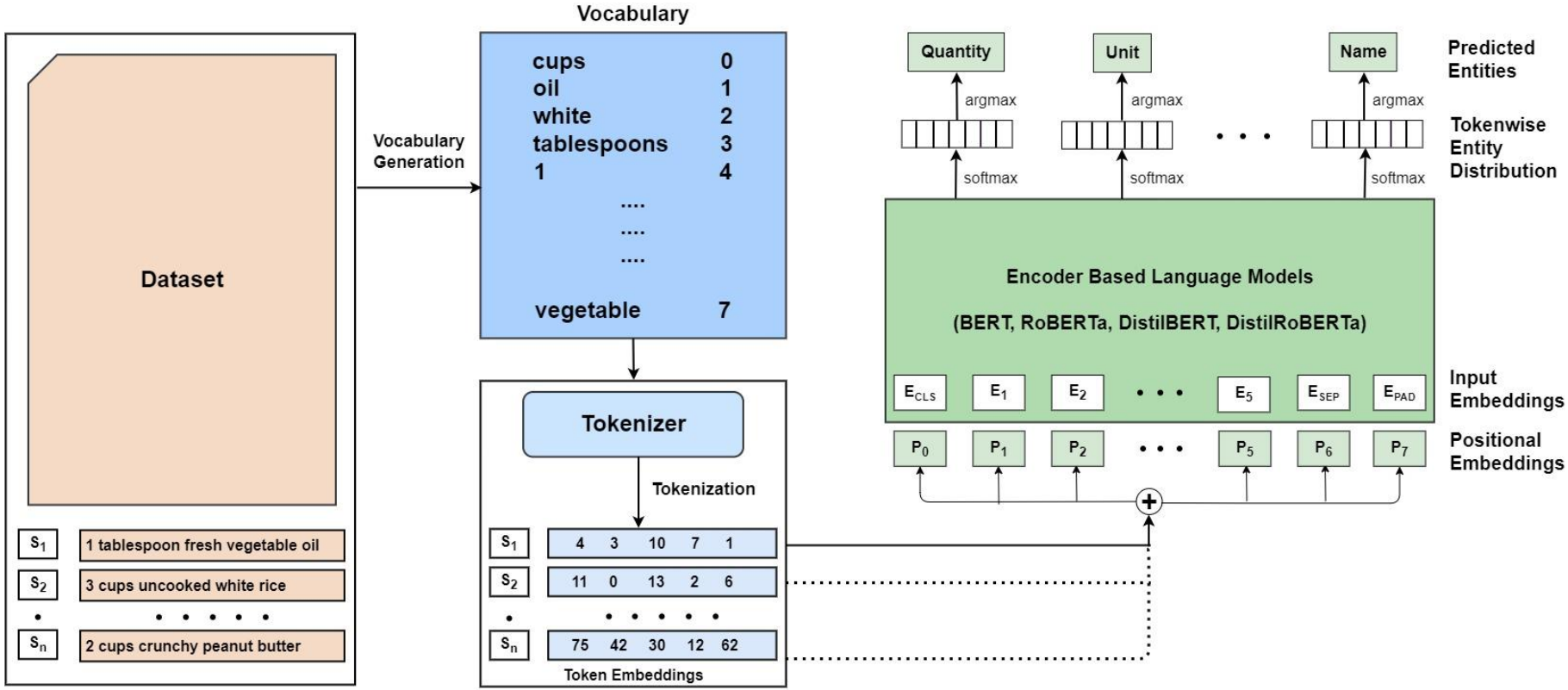
- The **salient contributions** of the research study are

a) **Creation** of augmented and machine-annotated ingredient phrase datasets

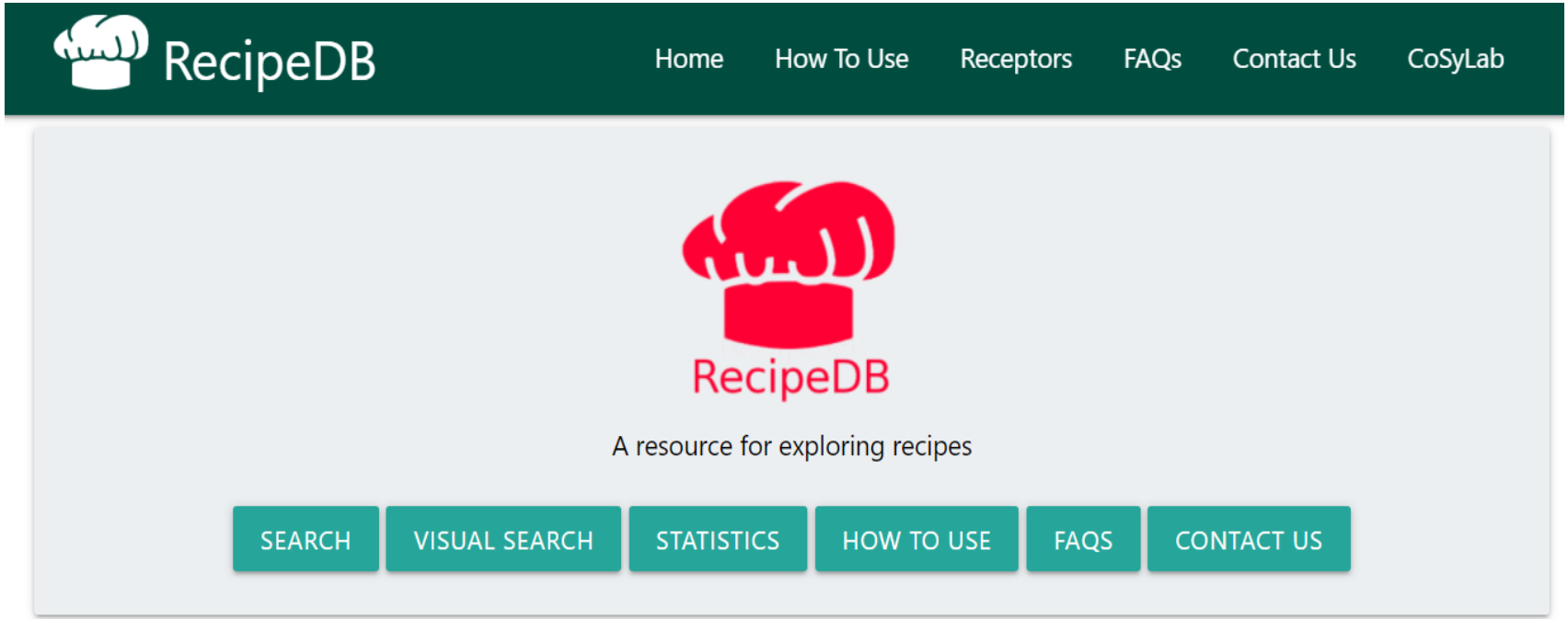
b) **Analysis** of the distribution of RecipeDB ingredient phrases

c) **Implementation of NER** approaches on recipe texts involving statistical, deep-learning-based fine-tuning of language models and few-shot prompting on LLMs.

Methodology



Dataset



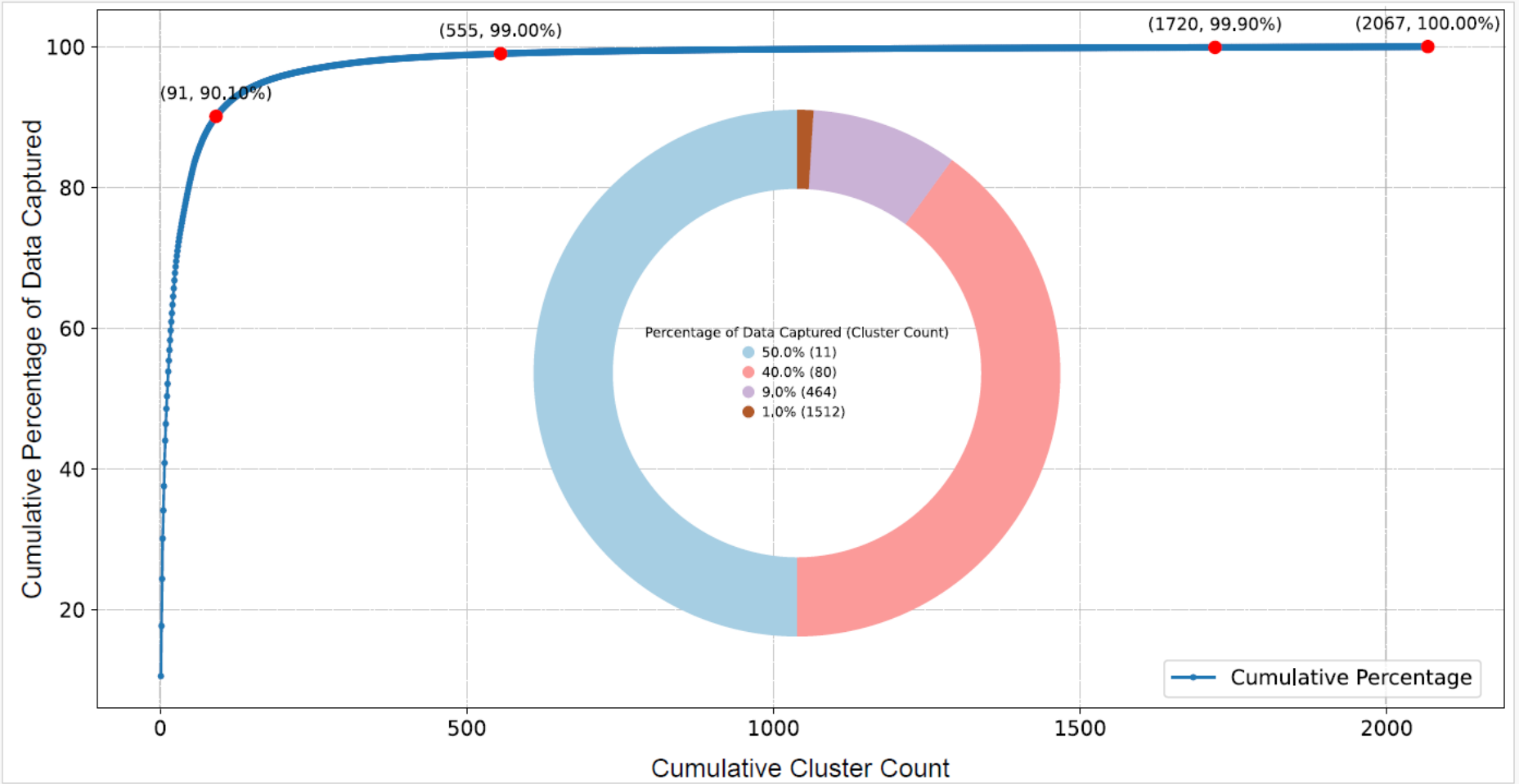
https://cosylab.iiitd.edu.in/recipedb/

Dataset: Augmented Dataset

	QUANTITY	UNIT	O	STATE	DF	NAME	NAME
Original	3	tablespoons	finely	chopped	fresh	ginger	root
LwTR	3	**cup**	finely	chopped	fresh	**onion**	root
SR	3	tablespoons	**superbly**	chopped	**new**	ginger	root
SiS	3	tablespoons	finely	chopped	fresh	**root**	**ginger**

Dataset: Machine-Annotated Dataset

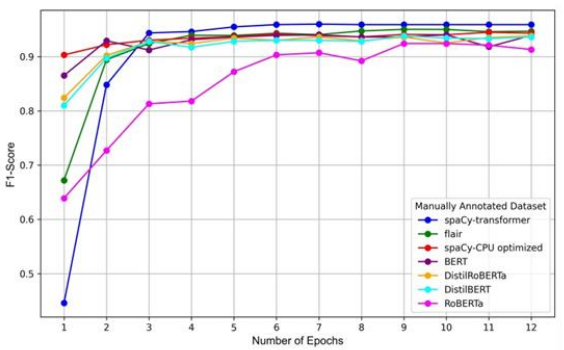
- Trained the **Stanford NER** on the labeled corpus (6,611 + 2,187 = 8,798 ingredient phrases) to annotate the unique ingredient phrases (349,762) from RecipeDB.
- Manually cleaned the machine-generated annotations to identify the error patterns and correct them programmatically.
- Implemented **Stratified Entity Frequency Sampling**, a clustering and sampling approach, which samples the dataset with varied ingredient phrase patterns.

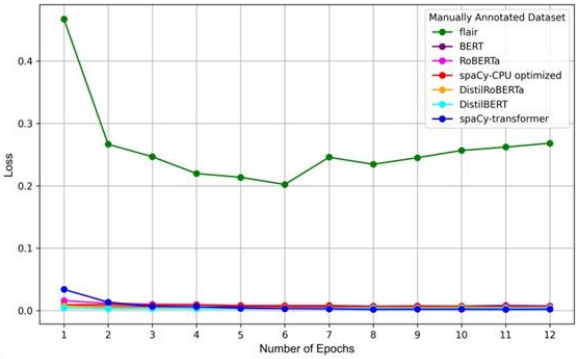

(91, 90.10%)
(555, 99.00%)
(1720, 99.90%)
(2067, 100.00%)
Cumulative Percentage of Data Captured
Cumulative Cluster Count
Cumulative Percentage
Percentage of Data Captured (Cluster Count)
50.0% (11)
40.0% (80)
9.0% (464)
1.0% (1512)

Modelling Techniques

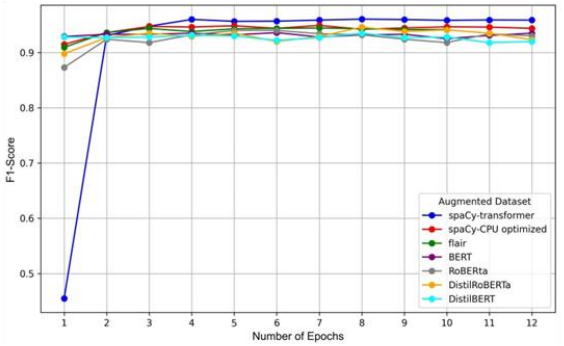
1. BERT [Toutanova et al. 2019]
2. DistilBERT [Sanh et al. 2019]
3. RoBERTa [Liu et al. 2023]
4. DistilRoBERTa [Sanh et al. 2019]
5. spaCy [Honnibal et al. 2020]
6. Flair [Akbik et al. 2019]
7. Stanford NER [Finkel et al. 2005]

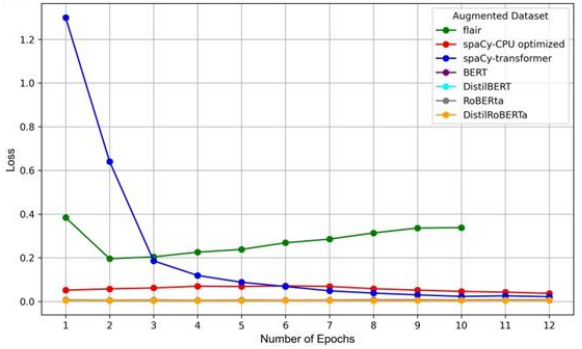
NER Model Comparison



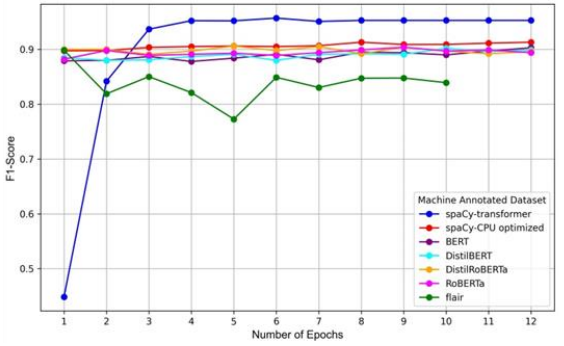


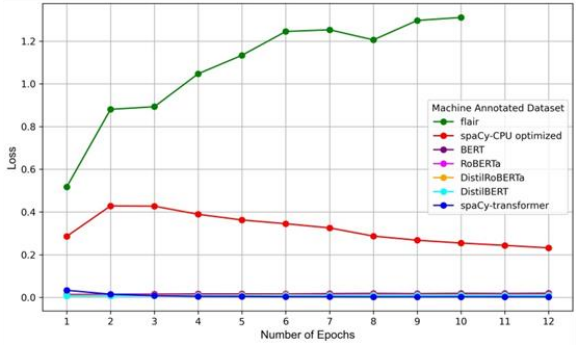
Manually Annotated





Augmented



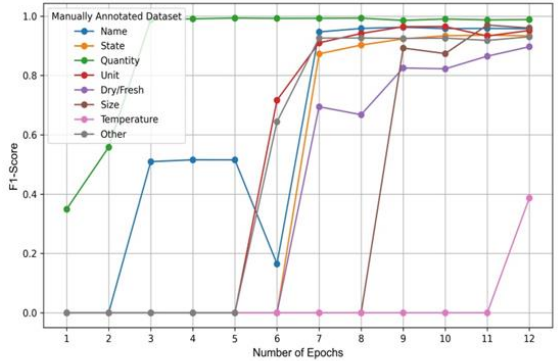


Machine Annotated

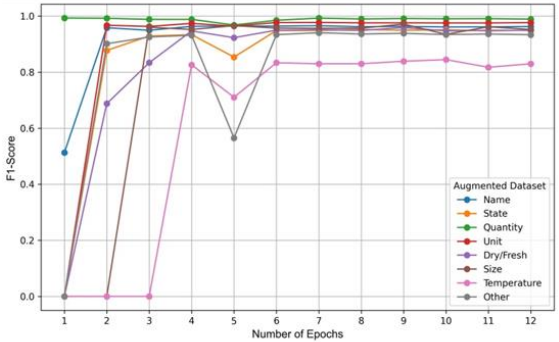
Results

Modelling Technique	Manually Annotated			Augmented			Machine Annotated		
	F1	P	R	F1	P	R	F1	P	R
spaCy-transformer	**95.90**	95.89	95.91	**96.04**	96.05	96.04	**95.71**	95.73	95.69
spaCy- CPU optimized	94.46	94.52	94.41	94.91	94.92	94.90	91.30	91.36	91.24
Stanford NER	95.52	95.64	95.39	95.16	94.37	95.96	89.9	91.31	88.53
DistilBERT	93.80	95.20	93.60	93.50	93.50	94.60	90.20	92.20	89.70
BERT	94.00	94.70	94.10	93.60	93.70	94.10	90.30	91.50	90.20
DistilRoBERTa	93.80	94.80	93.90	94.60	94.10	95.90	90.60	91.60	90.60
RoBERTa	92.40	92.90	92.60	94.00	94.50	94.10	90.40	91.60	90.20
flair	95.01	96.11	96.05	94.45	95.87	96.14	89.85	88.71	89.22

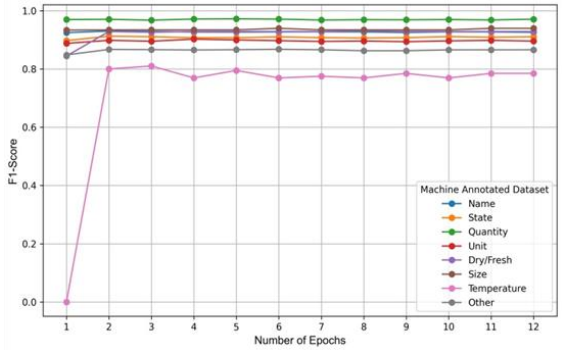
Tag-wise Analysis of Named Entities



Manually Annotated



Augmented



Machine Annotated

Analysis of Few-Shot Prompting on LLMs

Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.

Instruction:
Your task is to do Named Entity Recognition of input sentence. You must assign entity tags to each word in given input sentence from
[QUANTITY, UNIT, NAME, TEMP, STATE, SIZE, DF, O].

Number of tokens in input and output sentences must be equal.

Where,
NAME is the name of the ingredient added into the recipe, like onion, garlic etc.
UNIT is the unitary amount of the ingredient added into each step of recipe, like cup, tablespoon, etc.
QUANTITY is a multiple of the UNIT tag which gives the total quantity of the ingredient used in every step of the recipe.
TEMP is the temperature based state of the ingredient, like frozen, hot etc.
STATE is the condition of the ingredient used, like chopped, ground etc.
SIZE is the qualitative amount of the ingredient in each step of the recipe.
DF is the Dry or Fresh condition of the ingredient.
O is Others which is used for entities which are none of these : [QUANTITY, UNIT, NAME, TEMP, STATE, SIZE, DF]

Some Examples:

Input: '2 tablespoons vegetable oil , divided'
Output: [QUANTITY, UNIT, NAME, NAME, O, STATE]

Input: '2 tablespoons dried marjoram'
Output: [QUANTITY, UNIT, DF, NAME]

Input: '1 -LRB- 12 ounce -RRB- box Barilla Gluten Free Penne'
Output: [QUANTITY, O, QUANTITY, UNIT, O, UNIT, NAME, NAME, NAME, NAME]

Input: '2 jalapeno peppers , seeded and minced'
Output: [QUANTITY, NAME, NAME, O, STATE, O, STATE]

Input:
{input_sentence}
Output:

Results: NER using Few-Shot Prompting

Model	F1 Score
LLaMA2-7b	44.29
LLaMA2-13b	54.20
Mistral-7b	47.51
Vicuna-7b	51.41

Conclusions

- This study presents one of the most extensive **labeled data resources of named entities** from recipe ingredient phrases.
- We built deep-learning and statistical models to achieve **state-of-the-art results** in the domain of **culinary context**.
- Our present study focuses on only ingredient phrases while **not accounting for the recipe instructions**, which often carry semantic information about cooking that encodes cultural nuances.
- **In the future**, this research may be extended to include LLM fine-tuning, implementing NERs on cooking instruction, and implementation of multilingual NER.

References

1. Devansh Batra, Nirav Diwan, Utkarsh Upadhyay, Jushaan Singh Kalra, Tript Sharma, Aman Kumar Sharma, Dheeraj Khanna, Jaspreet Singh Marwah, Srilakshmi Kalathil, Navjot Singh, Rudraksh Tuwani, and Ganesh Bagler. 2020. RecipeDB: A resource for exploring recipes. Database, page 77.
2. Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In Association for Computational Linguistics, pages 54–59, Association for Computational Linguistics.
3. Gjorgjina Cenikj, Gasper Petelin, Barbara Korousic Seljak, and Tome Eftimov. 2022. SciFoodNER: Food Named Entity Recognition for Scientific Text. In Proceedings – IEEE International Conference on Big Data, pages 4065–4073. IEEE.
4. Gjorgjina Cenikj, Gorjan Popovski, Riste Stojanov, Barbara Koroušić Seljak, and Tome Eftimov. 2020. Butter: Bidirectional lstm for food named-entity recognition. In IEEE International Conference on Big Data (Big Data), pages 3550–3556. IEEE.

References

5. Nirav Diwan, Devansh Batra, and Ganesh Bagler. 2020. A named entity based approach to model recipes. Proceedings - 36th International Conference on Data Engineering Workshops, ICDEW, pages 88–93.
6. Mansi Goel and Ganesh Bagler. 2022. Computational gastronomy: A data science approach to food. Journal of Biosciences, 47(1):1–10.
7. Gorjan Popovski, Stefan Kochev, Barbara Koroušić Seljak, and Tome Eftimov. 2019. Foodie: A rule-based named-entity recognition method for food information extraction. ICPRAM 2019 - Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, 12:915–922.
8. Honnibal Matthew, Ines Montani, Sofie Van Landeghem, and Boyd Adriane. 2020. spaCy Industrial-strength Natural Language Processing in Python.
9. Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT.

Acknowledgments

Mansi Goel*, Ayush Agarwal*, Shubham Agrawal*, Janak Kapuriya*, Akhil Vamshi Konam*, Rishabh Gupta, Shrey Rastogi, Niharika and Ganesh Bagler.

Infosys Centre of Artificial Intelligence, Department of Computational Biology, and Department of Computer Science, IIIT-Delhi.

* Equal Contribution

Thank You

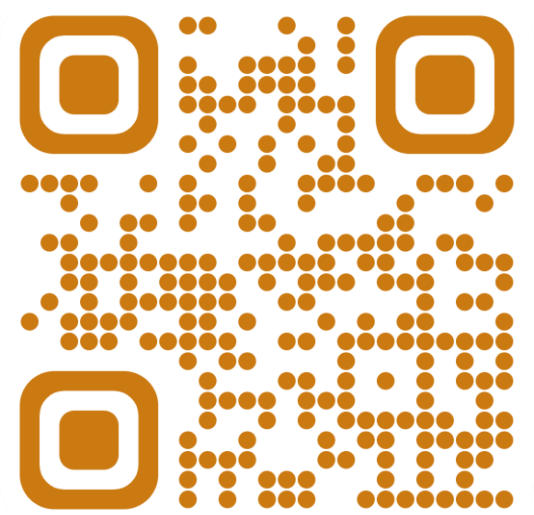
Any questions?

You can mail me at:

mansig@iiitd.ac.in

Complex Systems Laboratory, IIIT-Delhi

https://cosylab.iiitd.edu.in



Build Food-Tech Business with Computational Gastronomy APIs