Report No: 05

Task Title: Report on Research Presentation

Course Title: Mobile Application Development Laboratory
Course Code: CSE-410

4th Year, 1st Semester - 2023

Date of Performance: 17/09/2024

Date of Submission: 22/09/2024

Submitted to -

Dr. Md. Ezharul Islam	**Anup Majumder**	**Samsun Nahar Khandakar**
Professor	**Assistant Professor**	**Lecturer**

Submitted by –

Name	Roll	Exam Roll
Shanjida Alam	353	202165

Department of Computer Science & Engineering
Jahangirnagar University
Savar, Dhaka-1342

I have picked the three topics among 6 topics. They are: Retrieving Top K% Relevant Patterns for Distant Supervision-Based Relation Extraction for Bangla Sentences, Federated Learning and NLP for Advanced E-mail Forensics: A Privacy-Preserving Approach, Punctuation Prediction and Restoration for Bangla Text Using Transformer-based Learning.

First Topic: Retrieving Top K% Relevant Patterns for Distant Supervision-Based Relation Extraction for Bangla Sentences

Introduction:

Natural language processing (NLP) tasks such as relation extraction (RE) entail determining the relationships between items in a text.
Semantic annotations are a necessary tool for turning the rapidly expanding unstructured online information into structured data. Due to the number and diversity, manual annotation is not feasible, especially for low-resource languages like Bangla. Automating the creation of datasets by remote supervision with knowledge bases (such as DBpedia and Freebase) reduces noise by introducing assumptions about relations that are too simplistic.

Low-resource languages such as Bangla face a problem with relation extraction (RE) because it is difficult to manually annotate vast amounts of unstructured internet content into structured data because of its volume and complexity. Because it depends on extremely basic assumptions about the relationships between entities, distant supervision—which leverages knowledge bases such as DBpedia and Freebase to automate dataset creation—introduces noise, resulting in incorrect labels and decreased performance.

Proposed solution:

Selecting the top K% of valid patterns for each relation using conflict scores is the suggested method for mitigating the noise in distant supervision-based relation extraction (RE) for Bangla. By applying probability ratings to weed out noisy patterns, these patterns are further refined. The model then labels phrases with erroneous patterns as "None," which lowers noise and boosts precision. In order to improve the model's overall performance, many classifiers are combined and trained using an ensemble method.
They are used in three approaches in this proposed solution. These are:

1. **Pattern Selection:** Pattern selection involves identifying and choosing the top K% of the most accurate patterns (valid patterns) for each relation. These patterns are selected based on "conflict scores," which measure how often a pattern causes errors.
2. **Filtering:** Filtering means improving the selected patterns by using "probability scores" to assess their accuracy. Patterns that are likely to cause errors (noisy

patterns) are identified and relabeled as "NONE," meaning they are marked as irrelevant, so they don't affect the final results. This process helps ensure that only the most accurate patterns are used for relation extraction.

3. **Model Training:** Model training using an "ensemble method" means combining multiple models to improve overall performance. Instead of relying on just one model, different model are trained, and their results are combined (through methods like majority voting or averaging). This approach helps make the final predictions more accurate and reliable.

My opinion about this topic:

In my opinion, the approach of using distant supervision for relation extraction (RE) in Bangla is a smart solution for low-resource languages. Since manual annotation is not feasible for large datasets, this method helps automate the process, even though it introduces some noise. The proposed method of selecting and refining patterns through conflict and probability scores shows a smart way to tackle this noise, ensuring more accurate results. I especially like the use of combining multiple models to improve performance, which seems like an effective way to boost the reliability of relation extraction. Overall, this work is a great step forward in making NLP more accessible for lesser-known languages like Bangla.

Second Topic: Federated Learning and NLP for Advanced E-mail Forensics: A Privacy-Preserving Approach

Introduction:

The problem here is that email communication can sometimes show warning signs of security threats, like phishing. This happens when emails include unusual attachments, suspicious links, or when the language and tone don’t seem right. If someone asks for sensitive information unexpectedly, that’s often a big red flag for phishing attempts. To stay safe, it’s important to double-check any unexpected emails and be careful with attachments or links. Educating users and having strong security measures in place can really help reduce the chances of falling victim to these kinds of email-based attacks.

The main goals of this thesis are to create a system that analyzes email content using NLP (Natural Language Processing) to uncover important details. It also focuses on using federated learning with a method called TLBO to allow efficient, decentralized learning across different systems. Another aim is to ensure privacy while making the model’s decisions easy to understand, using tools like LIME. Overall, it’s all about improving the detection of complex threats in emails and making cybersecurity stronger.

Proposed Solution:

The proposed solution uses federated learning and Teaching-Learning-Based Optimization (TLBO) to improve email forensics. Here's how it works in simple terms:

1. **Federated Learning:** Instead of sending all data to a central server, each device (or client) trains its model locally on its data. The devices then share only the model updates with the central server, not the raw data, which helps maintain privacy.
2. **Teaching-Learning-Based Optimization (TLBO):** This method optimizes the models by having a "teacher" (the best-performing model) guide other models (students) to improve their learning process.

By combining these methods, the system can detect email threats more accurately while keeping sensitive data private. Additionally, Explainable AI (like LIME) is used to make the decisions made by the system more transparent and easier to understand. Overall, this approach makes email security more efficient and trustworthy.

My opinion about this topic:

In my opinion, the idea of combining federated learning with NLP for email forensics is a really smart way to improve privacy and security. It’s cool that instead of sending sensitive data to a central server, the learning happens on each device, keeping the data safe. Adding the Teaching-Learning-Based Optimization (TLBO) approach to guide the models makes the whole system even more efficient. Also, making the decisions understandable with tools like LIME adds an extra layer of transparency. Overall, I think this approach is a great way to tackle cybersecurity challenges while respecting user privacy.

Third Topic: Punctuation Prediction and Restoration for Bangla Text Using Transformer-based Learning

Introduction:

The problem we are dealing with is that Bangla text often lacks proper punctuation, which makes it harder to read and understand. Punctuation like commas, periods, and question marks play a big role in guiding the reader through the text. Without these marks, sentences can become confusing or change their meaning entirely.

In comparison to high-resource languages like English, Bangla doesn't have as many automated tools to help with adding punctuation. This is partly due to the lack of large, high-quality datasets and the complexity of Bangla's script and grammar. So, there’s a clear need for better AI systems to handle punctuation prediction in Bangla, which will make Bangla text more readable and help improve AI applications for the language.

Currently, AI tools don't work well for the Bangla language because of several reasons:

- There's a lack of large, high-quality datasets.
- Bangla has many dialects and regional variations, yet there's little research being done to develop AI tools to address these.
- The Bangla script and grammar are quite complex, making it harder to build efficient AI solutions.

So, there's an urgent need to develop AI systems that are specifically tailored for Bangla.

Proposed Solution:

The proposed solution for punctuation prediction and restoration in Bangla text uses a transformer-based learning model, specifically BERT. Here's a short and simple explanation:

The solution involves training the BERT model to automatically add punctuation (like commas, periods, and question marks) to Bangla text. The model is trained on a dataset collected from sources like Wikipedia and newspapers. Each word or character is labeled based on its punctuation, and the model learns to predict the correct punctuation in the output. This approach improves the readability of machine-generated Bangla text by making it more similar to human-written text. It also helps bridge the gap between existing AI tools and the complex nature of the Bangla language.

My opinion about this topic:

In my opinion, the topic of punctuation prediction for Bangla text is really interesting and much-needed. Since Bangla is a low-resource language, it doesn’t have the same level of automated tools as English. This makes it harder for students and professionals working in Bangla to create clean and structured text. I like the idea of using AI, especially models like BERT, to bridge this gap. It not only helps improve the readability of text but also makes sure machine-generated content is easier to understand. Overall, I think it’s a great initiative to make Bangla more accessible in the digital space.