IITP-CUNI@3C:Supervised Approaches for Citation Classification (Task A) and Citation Significance Detection (Task B)

Kamal Kaushik Varanasi†, Tirthankar Ghosal‡, Piyush Tiwary† and Muskaan Singh‡

†Indian Institute of Technology Patna, India, ‡ Charles University, Czech Republic

3C Shared Task

- **Task A (Citation Context Classification Based on Purpose) :**
  - Multiclass Classification Problem (6 labels).
  - Labels : BACKGROUND, USES, COMPARES CONTRASTS, MOTIVATION, EXTENSION, and FUTURE.
- **Task B (Citation Context Classification Based on Influence) :**
  - Binary Classification Problem.
  - Labels : INCIDENTAL or INFLUENTIAL

Approach for Task A

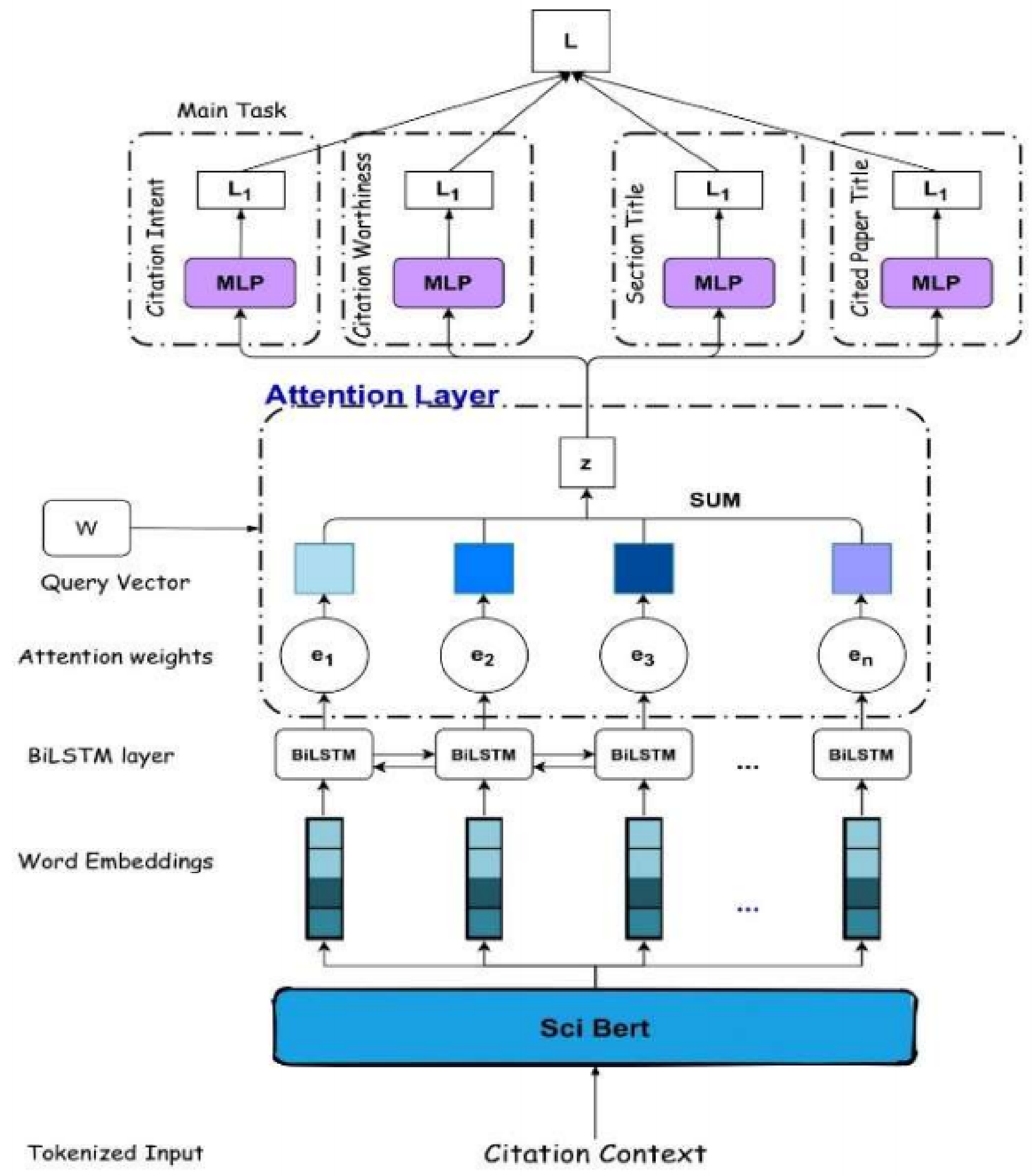
We use a Multi-Task Learning Framework that incorporates three scaffold tasks. Two of them are the Structural scaffolds (section title and citation worthiness) inspired by work done in Cohan et al. (2019) that help in leveraging the relationship between the structure of the research papers and the intent of the citations. The third scaffold is the Cited paper title scaffold.

- **Section Title Scaffold (S1):**
  - Predicts section under which the citation occurs.
  - Researchers follow a standard order while presenting their scientific work in the form of sections.
  - So, citations may have different nature according to the section under which they are cited.
  - Example : Results-comparison related citations are often cited under the Results section.
- **Citation Worthiness Scaffold (S2):**
  - Predicts whether a sentence needs a citation or not. Or Classify whether a sentence is citation text or not.
  - Writing style of citation sentences is different than the normal sentences.
- **Cited Paper Title Scaffold (S3):**
  - Leverages relation between citation context and the cited paper.
  - Sometimes a citation context might be ambiguous, making it difficult to predict the intent of the citation correctly.
  - Additional context from the cited paper (abstract, title, etc) can help in such cases.
  - Input : Concatenated vector of citation context and the cited paper title fields from the 3C train data.

Code

`https://github.com/vkk1710/IITP-NAACL-SDP-3C-Shared-Task`

Task A : Model Architecture



Training

- **Training on SciCite :**
  - We only train the Main task and the Structural scaffolds.
  - Note that SciCite contains the data for the structural scaffolds.
- **Fine Tuning on 3C Task A Train Data :**
  - We fine tune on the Main Task and Cited paper title scaffold, while freezing the parameters of other scaffolds.

Approach for Task B

We pursue a feature-engineering approach to curate simple features from cited-citing paper pairs. We use traditional machine learning algorithms (SVM, KNN, Decision Tree, Random Forest and XGBoost) to classify on the basis of the extracted features. We choose our best model (Random Forest) on the basis of the performance on the validation data. The features extracted are given below.

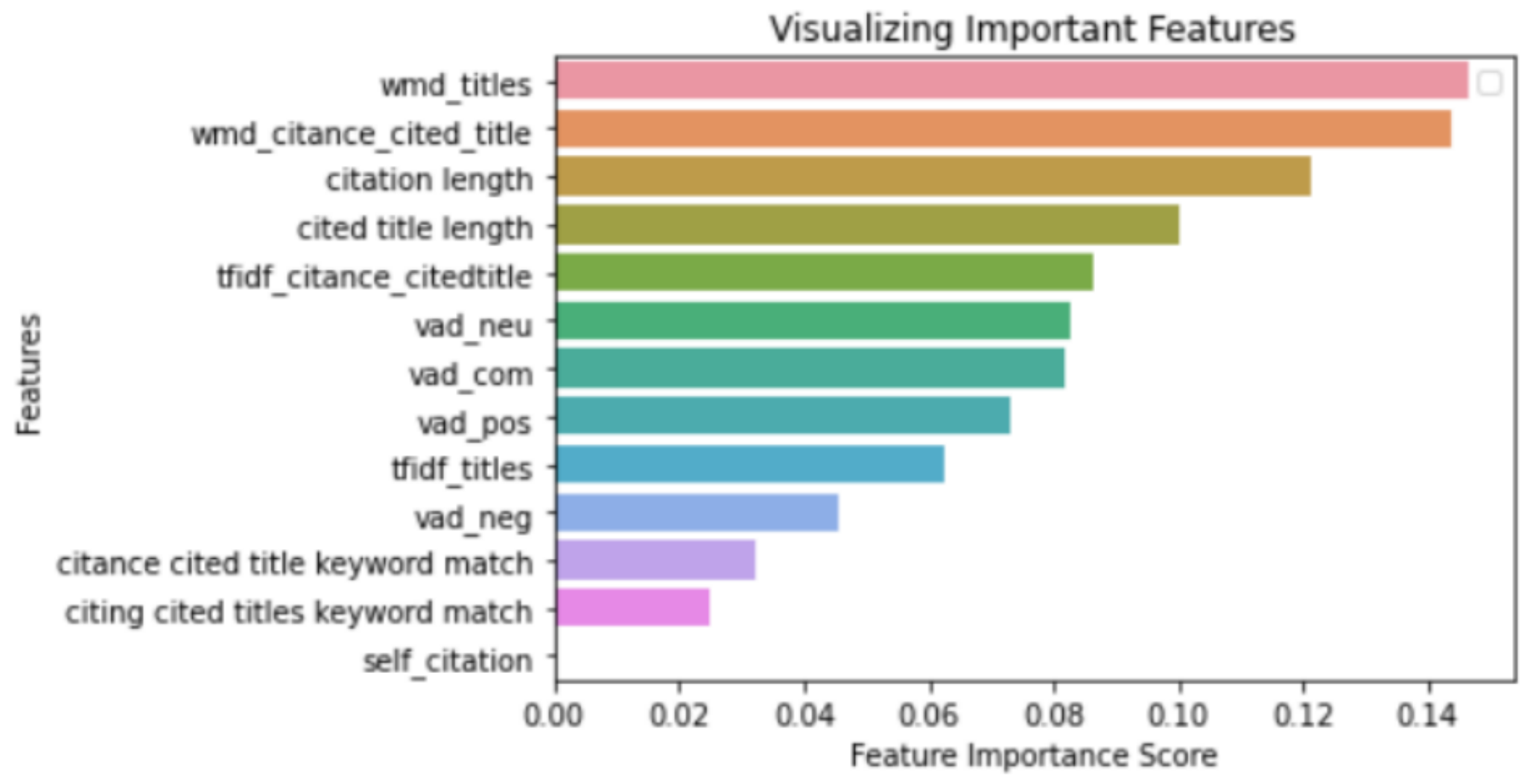
- **tf-idf Features :**
  - We measure the cosine similarity between the tf-idf representations of the 1. Titles of cited and citing papers and 2. Citation context and the Title of the cited paper.
  - Titles of Cited papers may contain information regarding their contribution or purpose of the paper.
  - Hence higher lexical similarity with Citation context may construe that the cited paper may have been used significantly in the current paper.
- **Word Mover's Distance Features :**
  - We measure similarity among pairs of Citation Context, Titles of Citing and Cited papers in semantic space.
- **VADER Polarity Index Positive, Negative, Neutral, Compound :**
  - We measure the VADER polarity index to quantify the intensity of the positive/negative emotion of the citation context.
- **Keyword Overlap :**
  - We compare the number of common keywords between 1. Title of citing and cited paper and 2. Citance and the title of cited paper.
- **Length Features :**
  - We measure the length of Citation Context and Title of Cited paper.
  - More the number of words spent by Citing paper on Cited paper, more significance of the Cited paper.
- **Self Citation :**
  - We check if the authors of the citing and cited paper are the same.
  - This might be the case of self-citation or can also signal the extension of the work.

Feature Importance

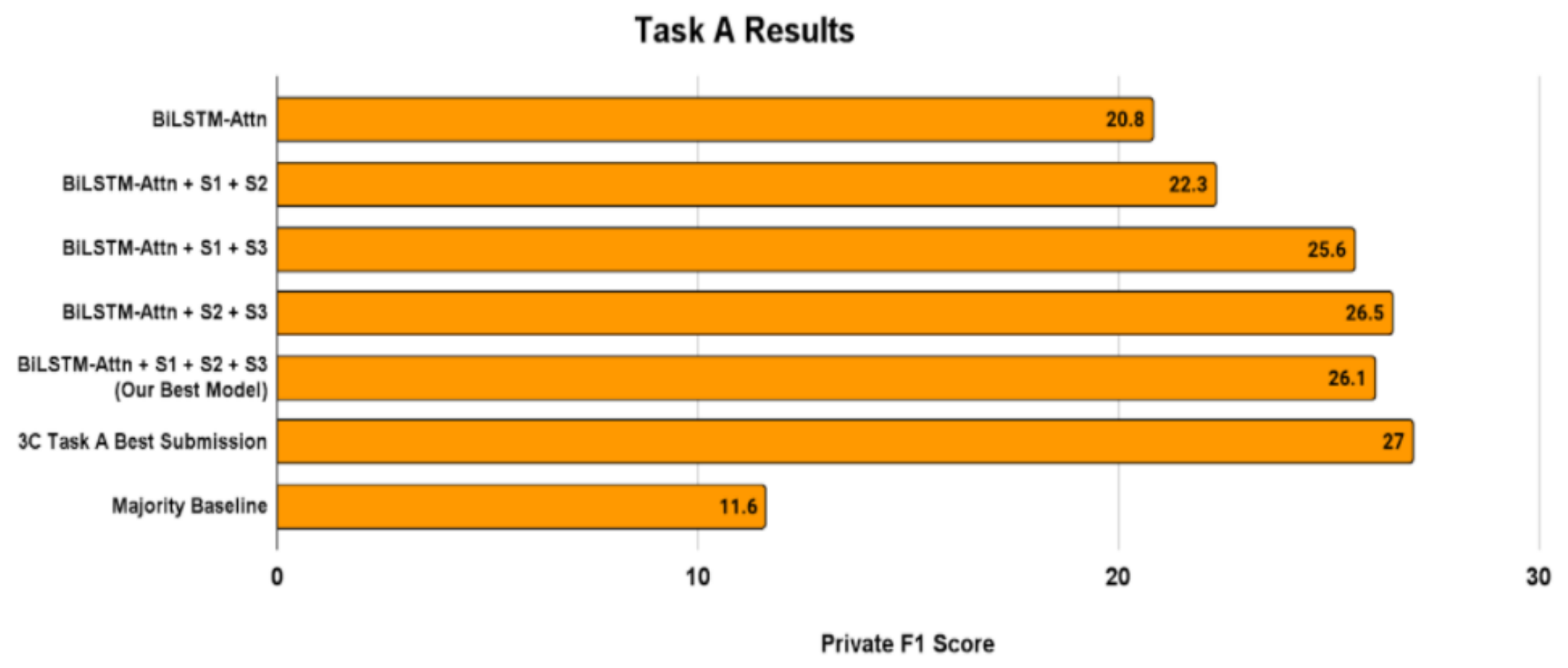
- We also compute the feature significance graph. It is evident that the semantic similarity features (WMD features) are more significant than the lexical similarity (tf-idf) and sentiment intensity (VADER) features.

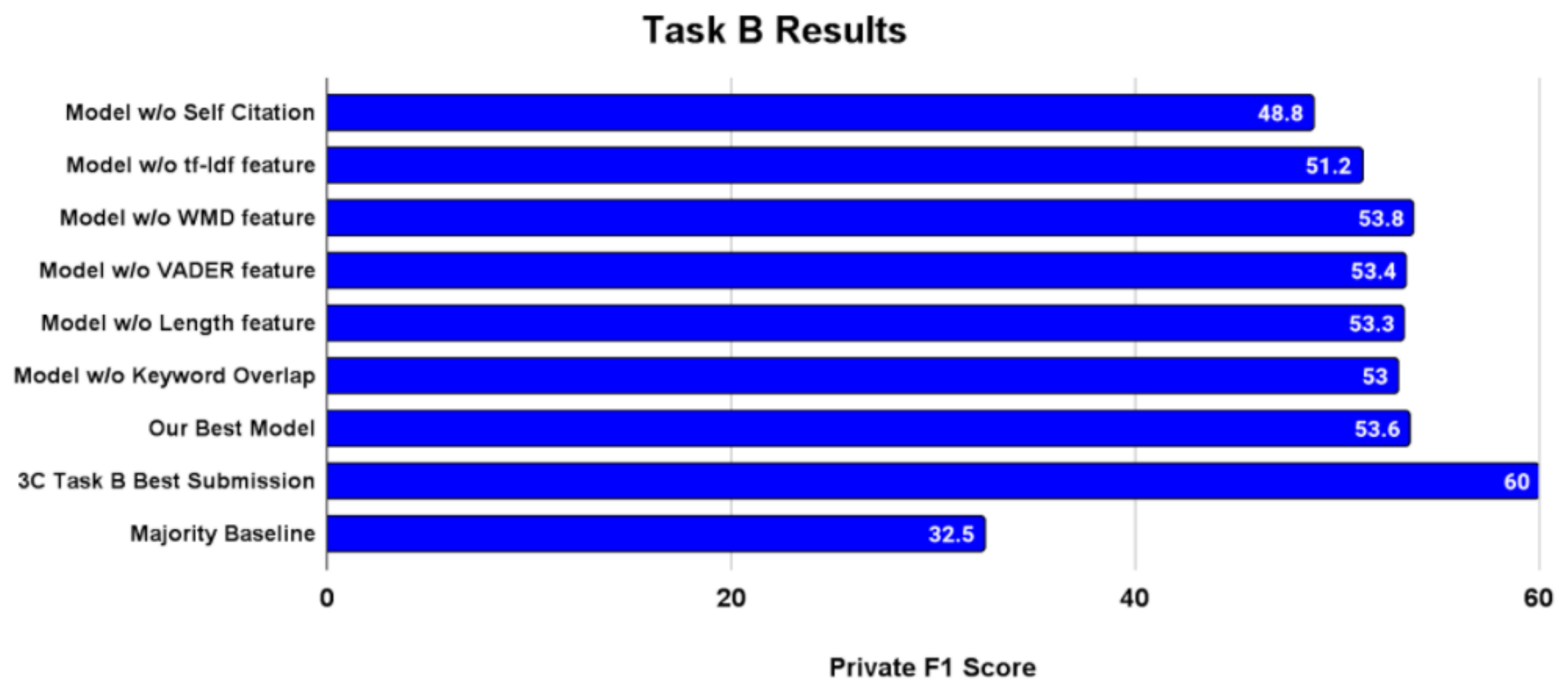
Feature Importance

- This trend might be true because in general, research articles have a style of writing that involves significantly less subjective content and follows a more objective discourse.



Results





- For Task A, we achieve comparable results with respect to the best performing system in the competition.
- In case of Task B, we achieve better performance than the majority baseline by using some simple features.

Future Work

- Use Abstracts and Full text information of cited-citing paper pairs as additional context for both the tasks.
- Solve the problem of overfitting on the given data.