Information Retrieval and Extraction

Final Semester Examination

Three Sections **180 Minutes** **180 Marks**

Section I: Short Answer Questions (Answer ANY SIX questions – 60 Marks)

1. What does attention mean in context of transformer networks? Explain the differences between cross attention and self-attention.
2. What are the desirable characteristics of a search engine? Describe the various measures used for evaluating a retrieval system in detail.
3. Describe the workings of the page rank algorithm in brief and mention its merits and demerits.
4. Explain the difference between WordNet and GloVe. Provide examples of use cases where one might be more appropriate than the other.
5. Describe three different techniques for performing NER (Named Entity Recognition). Also mention the pros and cons of each of those.
6. What are the key differences between BERT and GPT? In which scenarios would you prefer one over the other?
7. Explain the ROUGE evaluation metrics and their variations. Devise an alternative or modified metric which can also capture the semantic information during evaluation.

Section II: Project problems (Answer 3 questions – 60 Marks)

Do NOT answer your own major project related question.

1. Reference project: Movie ratings and Recommendations

In this Project, the team aimed at predicting the ratings of a movie using movie metadata along with plot summary. On top of that, given a set of user ratings and a list of movies, they also tried to predict a user specific rating for the movie.

Assuming you are free to use the resources publicly available on the internet (ex IMDb, Rotten Tomatoes etc.). What features would you use apart from the ones mentioned by the team, in order to have a more efficient system? Also note that noise in the input can even harm the performance of the model, hence choose the features carefully and explain their relevance.

For this question, please describe the features, why you think they are relevant and how you would incorporate them in the pipeline. Feel free to describe the aspects of the pipeline which would benefit from your selected features.

2. Reference project: Multilingual Tweet intimacy Analysis

The objective of the project was to devise a supervised pipeline to provide an intimacy score (between 0 to 5) for tweets in 6 different languages.

How would you devise an improved modelling technique for this problem? Please mention details like choice of the loss function, model architecture, evaluation methods etc. Also mention the reasoning behind the choices

3. Reference project: Code comment generation and classification

The aim of the project was to create a classifier to classify the comments as relevant or irrelevant, followed by a pipeline to automatically generate code comments and summaries

Describe in detail the possible applications and extensions of this work. For each application or extension listed, discuss the technical feasibility. Also list the steps required to adapt this solution for the applications mentioned.

ASTs → SiT Bast Adamo

4. Reference Project: EVALUATING TRANSFORMER MODELS PERFORMANCE ON REAL CLAIMS DATA

The project focused on answering two prominent research questions:

Can the models (Transformer based) trained and tested on FEVER / FEVEROUS (Fact Extraction and Verification) dataset, perform well on 'real' datasets. Does finetuning on real dataset, improve performance

Is there any bias in verdict while fake claims are fact checked? Is the verdict 'accurate' based on evidence researched and published by Fact Checkers?

With respect to the analysis done in this project and other examples from the content of the course, comment on the feasibility or infeasibility of adapting deep learning-based solutions in the real world.

Also discuss the reasons why a transformer-based model might be prone to bias of various kinds and the possible ways of mitigating it.

Section III: Research Paper (60 Marks)

Read the attached short research paper on "Task Adaptive Pretraining of Transformers for Hostility Detection" and answer the following questions:

1. Write a 350-word summary of the main contributions of the paper in terms of problem definition, solution outline, and evaluation.

2. What are the IR/IE specific contributions if any? What more IR specific contributions or approaches can you think of for this task?

3. What are the Machine Learning specific contributions, what insights do they present?

4. What are the limitations of the approach?

5. What improvements can be made to the evaluation or approach described in the paper?

6. What are the possible uses/applications of this work? Try to list at least five with short descriptions.

Task Adaptive Pretraining of Transformers for Hostility Detection

Tathagata Raha and **Sayar Ghosh Roy** and **Ujwal Narayan** and **Zubair Abid** and **Vasudeva Varma**
Information Retrieval and Extraction Lab
International Institute of Information Technology, Hyderabad, India

Abstract

Identifying adverse and hostile content on the web and, more particularly, social media has become a problem of paramount interest in recent years. With their ever-increasing popularity, fine-tuning of pretrained Transformer-based encoder models with a classifier head is gradually becoming the new baseline for natural language classification tasks. Our work explores the gains attributed to Task Adaptive Pretraining (TAPT) before fine-tuning of Transformer-based architectures. We specifically study two problems, namely, (a) Coarse binary classification of Hindi Tweets into Hostile or Not, and (b) Fine-grained multi-label classification of Tweets into four categories: hate, fake, offensive, and defamation. Building upon an architecture that takes emojis and segmented hashtags into consideration for classification, we were able to showcase the performance upgrades due to TAPT experimentally. Our system (with team name 'iREL IIIT') ranked first in the 'Hostile Post Detection in Hindi' shared task with an F1 score of 97.16% for coarse-grained detection and a weighted F1 score of 62.96% for fine-grained multi-label classification on the provided blind test corpora.

1 Introduction

With the increase in the number of active users on the internet, the amount of content available on the World Wide Web, and more specifically, that on social media, has seen a sharp rise in recent years. A sizable portion of the available content contains hostility, thereby posing potential adverse effects upon its readers. Content that is hostile in the form of, say, a hateful comment, unwarranted usage of offensive language, attempt at defaming an individual, or a post spreading some misinformation circulates faster as compared to typical textual information (Mathew et al., 2019; Vosoughi et al., 2018). Identifying and pinpointing such instances of hostility is of utmost importance for ensuring the sanctity of the World Wide Web and the well-being of its users. And as such, multiple endeavors have been made to design systems that can automatically identify harmful content on the web (Badjatiya et al., 2019; Pinnaparaju et al., 2020; Kumar et al., 2018; Badjatiya et al., 2017; Mandl et al., 2020).

In this work, we focus on the problem of identifying specific Hindi Tweets which are hostile. We further analyze whether the Tweet can fit into one or more of the following buckets: hateful, offensive, defamation, and fake. The popularity of pretrained Transformer-based (Vaswani et al., 2017) models for tasks involving Natural Language Understanding is slowly making them the new baseline for text classification tasks. In such a scene, we experiment with Task Adaptive Pretraining (Gururangan et al., 2020). IndicBERT (Kakwani et al., 2020), which is similar to BERT (Devlin et al., 2018) but trained on large corpora of Indian Language text is our primary pretrained Transformer of choice for dealing with Hindi text.

Table 1: Distribution of Supervised labels in Training set

Label	Frequency
Non-Hostile	3050
Defamation	564
Fake	1144
Hate	792
Offensive	742

Table 2: Distribution of labels in the Test set

Label	Frequency
Non-Hostile	873
Defamation	169
Fake	334
Hate	234
Offensive	219

We adopt a model architecture similar to Ghosh Roy et al., 2021 (Ghosh Roy et al., 2021), which leverages information from emojis and hashtags within the Tweet in addition to the cleaned natural language text. We are able to portray 1.35% and 1.40% increases for binary hostility detection and, on average, 4.06% and 1.05% increases for fine-grained classifications into the four hostile classes on macro and weighted F1 metrics respectively using Task Adaptive Pretraining (TAPT) before fine-tuning our architectures for classification.

The organizers of the Constraint shared task[1] provided the dataset for training and model development (Bhardwaj et al., 2020; Patwa et al., 2021). The data was in the form of Tweets primarily composed in the Hindi language and contained annotations for five separate fields. Firstly, a coarse-grained label for whether the post is hostile or not was available. If a Tweet were indeed hostile, it would not carry the 'not-hostile' tag. Hostile Tweets carried one or more tags indicating its class of hostility among the following four non-disjoint sets (the Shared Task organizers provided the definitions for each category):

1. **Fake News**: A claim or information that is verified to be untrue.

[1] constraint-shared-task-2021.github.io

2. **Hate Speech**: A post targeting a specific group of people based on their ethnicity, religious beliefs, geographical belonging, race, etc., maliciously intends to spread hate or encourage violence.

3. **Offensive**: A post containing profanity, impolite, rude, or vulgar language to insult a targeted individual or group.

4. **Defamation**: A misinformation regarding an individual or group.

A collection of 5728 supervised training examples were provided, which we split into training and validation sets in an 80-20 ratio, while a set of 1653 Tweets served as the blind test corpora. The mapping from a particular class to its number of training examples has been outlined in Table 1. The distribution of labels within the test set is shown in Table 2. Note that the test labels were released after the conclusion of the shared task. Throughout, a post marked as 'not-hostile' cannot have any other label while the remaining posts can theoretically have n labelings, $n \in \{1, 2, 3, 4\}$.

In this section, we describe our model in detail and present the foundations for our experiments. We acknowledge that the language style for online social media text differs from that of formal and day-to-day spoken language. Thus, a model whose input is in the form of Tweets should be aware of and leverage information encoded in the form of emojis and hashtags. We base our primary architecture on that of Ghosh Roy et al., 2021 (Ghosh Roy et al., 2021) with a few modifications.

1.1 Preprocessing and Feature Extraction

Similar to Ghosh Roy et al., 2021 (Ghosh Roy et al., 2021), the raw input text is tokenized on whitespaces plus symbols such as commas, colons, and semicolons. All emojis and hashtags are extracted into two separate stores. The cleaned Tweet text, our model's primary information source, is free from non-textual tokens, including smileys, URLs, mentions, numbers, reserved words, hashtags, and emojis. The tweet-preprocessor[2] python library was used for categorizing tokens into the classes mentioned above.

To generate centralized representations of all emojis, we utilize emoji2vec (Eisner et al., 2016) to generate 300-dimension vectors for each emoji and consider the arithmetic mean of all such vectors. We use the ekphrasis[3] Python library for hashtag segmentation. The segmented hashtags are arranged in a sequential manner separated by whitespaces, and this serves as the composite hashtag or 'hashtag flow' feature. Thus, we leverage a set of three features, namely, (a) the cleaned textual information, (b) the collective hashtag flow information, and (c) the centralized emoji embedding.

1.2 Architecture

This subsection outlines the flow of information pieces from the set of input features to label generation. We leverage two Transformer models to generate embeddings of size 768 for the cleaned text and hashtag flow features. The two Transformer-based embeddings are passed through two linear layers to yield the final vector representations for cleaned text and hashtag collection. The three vectors: cleaned text, composite hashtag, and centralized emoji representation are then concatenated and passed through a linear layer to form the final 1836-dimension vector used for classification. A dense multi-layer perceptron serves as the final binary classifier head. The overall information flow is presented in Figure 1. For the multi-label classification task, we trained our architecture individually to yield four separate binary classification models. In all cases, we performed end-to-end training on the available training data based on cross-entropy loss.

[2] github.com/s/preprocessor
[3] github.com/cbaziotis/ekphrasis

Table 3: Results on the Validation split for every category (% Weighted F1 Scores)

Metric	Without TAPT	With TAPT	Gains
Hostility (Coarse)	96.87	98.27	1.40
Defamation	86.47	86.31	-0.16
Fake	89.53	90.99	1.46
Hate	85.69	87.06	1.37
Offensive	87.12	88.66	1.54

Table 4: Results on the Validation split for every category (% Macro F1 Scores)

Metric	Without TAPT	With TAPT	Gains
Hostility (Coarse)	96.84	98.19	1.35
Defamation	59.43	63.38	3.95
Fake	83.69	86.52	2.83
Hate	70.77	74.20	3.43
Offensive	68.72	74.73	6.01

1.3 Task Adaptive Pretraining

We turn to Gururangan et al., 2020 (Gururangan et al., 2020), which showcases the boons of continued pretraining of Transformer models on natural language data specific to particular domains (Domain Adaptive Pretraining) and on the consolidated unlabelled task-specific data (Task Adaptive Pretraining). Their findings highlighted the benefits of Task Adaptive Pretraining (TAPT) of already pretrained Transformer models such as BERT on downstream tasks like text classification. We experimented with the same approach for our task of hostility detection in Hindi, having IndicBERT as our base Transformer model. Our results (in section ??) showcase the gains attributed to this continued pretraining with masked language modeling (MLM) objective. Note that only the cleaned text encoder Transformer is undergoing TAPT. The hashtag sequence encoder Transformer is initialized to pretrained IndicBERT weights. We create a body of text using all of the available training samples, and in that, we add each sample twice: firstly, we consider it as is, i.e., the raw Tweet is utilized, and secondly, we add the cleaned Tweet text. A pretrained IndicBERT Transformer is further pretrained upon this body of text with the MLM objective. We use these Transformer model weights for our cleaned text encoder before fine-tuning our complete architecture on the labeled training samples.

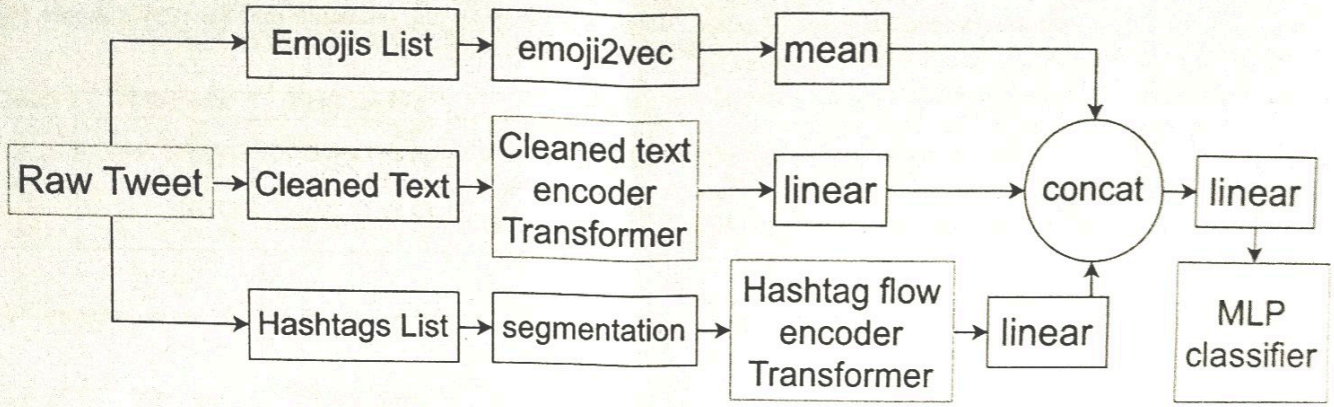

Figure 1: Model Architecture

Table 5: Shared Task Results: Top 3 teams on public leaderboard (% F1 Scores)

Metric	iREL IIIT (Us)	Albatross	Quark
Hostility (Coarse)	**97.16**	97.10	96.91
Defamation	**44.65**	42.80	30.61
Fake	77.18	**81.40**	79.15
Hate	**59.78**	49.69	42.82
Offensive	**58.80**	56.49	56.99
Weighted (Fine)	**62.96**	61.11	56.60

2 Results and Discussion

In Tables 3 and 4, we present metrics computed on our validation set. We observe 1.35% and 1.40% increases in the macro and weighted F1 scores for binary hostility detection and, on average, 4.06% and 1.05% increases in macro and weighted F1 values for fine-grained classifications into the four hostile classes. In all classes (except for 'Defamation' where a 0.16% performance drop is seen for the Weighted F1 metric), the classifier performance is enhanced upon introducing the Task Adaptive Pretraining. In Table 5, we present our official results with team name 'iREL IIIT' on the blind test corpora and compare it to the first and second runner-ups of the shared task.

3 Conclusion

In this paper, we have presented a state-of-the-art hostility detection system for Hindi Tweets. Our model architecture utilizing IndicBERT as the primary Transformer encoder, which is aware of features relevant to online social media style of text in addition to clean textual information, is capable of both identifying hostility within Tweets and performing a fine-grained multi-label classification to place them into the buckets of hateful, defamation, offensive, and fake. Our studies proved the efficacy of performing Task Adaptive Pretraining (TAPT) of Transformers before using such encoders as components of a to-be fine-tuned architecture. We experimentally showed 1.35% and 1.40% gains for coarse hostility detection and average gains of 4.06% and 1.05% for the four types of binary classifications, on macro and weighted F1 score metrics, respectively, in both cases. Our system ranked first in the 'Hostile Post Detection in Hindi' shared task with an F1 score of 97.16% for coarse-grained detection and a weighted F1 score of 62.96% for fine-grained classification on the provided blind test corpora.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, WWW '19, page 49–59, New York, NY, USA. Association for Computing Machinery.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Mohit Bhardwaj, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2020. Hostility detection dataset in hindi.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bosnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *CoRR*, abs/1609.08359.

Sayar Ghosh Roy, Ujwal Narayan, Tathagata Raha, Zubair Abid, and Vasudeva Varma. 2021. Leveraging multilingual transformers for hate speech detection. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*. CEUR.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

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 EMNLP*.

Ritesh Kumar, Atul Kr. Ojha, Marcos Zampieri, and Shervin Malmasi, editors. 2018. *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages). In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*. CEUR.