IRE Major Projects

Project Number: 1
Project Title: Emotion intensity prediction independent of the emotion labels.
Mentor(s): Himanshu (email: himanshu.maheshwari@research.iiit.ac.in)
Details:

1. This project has two sub problems:
    a. Students are required to predict the emotional intensity of the text. The training and testing data contains only one emotion (i.e. if training data contains anger, testing data also contains anger).
    b. Students are required to predict the emotional intensity of the text. The training data and testing data may contain text from multiple emotions. For example, the training data may have text from anger, fear and joy while testing data is only on sadness. The emotions that occur in the training phase do not occur in the testing phase, thus the system has to be robust of the emotion label.
2. Resource: SemEval 2018

Project Number: 2
Project Title: Collecting domain specific tables from Wikipedia articles
Mentor(s): Gaurang, Bhavyajeet, Daksh, Vasu & Harsha (email: harsha.pamidipalli@research.iiit.ac.in, daksh.rawat@students.iiit.ac.in)
Details:

1. Each group works on one specific domain (**Total 4 teams will be allocated this project - domain is different for each team**). Mentor(s) will allocate the domain for each team.
2. We will provide well defined domain/subdomain specification, each team uses data collection and preprocessing techniques to collect the data from the tables available on Wikipedia.
3. Augment this data with other information available within the Wikipedia ecosystem.
4. Enhance this data further by adding information outside the Wikipedia ecosystem from structured and unstructured sources.
5. The output is expected to be a JSON file with clean and high quality data.
6. Quantity of the data collected is an evaluation parameter.

Project Number: 3
Project Title: Collecting domain specific data given a seed set of structured and unstructured sources.
Mentor(s): Gaurang, Bhavyajeet, Daksh, Vasu & Harsha (email: harsha.pamidipalli@research.iiit.ac.in, daksh.rawat@students.iiit.ac.in)
Details:

1. Each group works on one specific domain (**Total 4 teams will be allocated this project** - domain is different for each team). Mentor(s) will allocate the domain for each team.
2. We will provide well defined domain/subdomain specification, each team uses data collection and preprocessing techniques to collect the data.
3. Augment this data with other information available within the Wikipedia ecosystem.
4. The output is expected to be a JSON file with clean and high quality data.
5. Quantity of the data collected is an evaluation parameter.

Project Number: 4
Project Title: Multihop document question generation
Mentor(s): Tanmay Sachan (email: tanmay.sachan@research.iiit.ac.in)
Details:

1. The task requires students to find an approach towards generating deep questions that require reasoning from multiple documents.
2. The dataset is well defined (HotpotQA).
3. As a starting point (using simple models), students will be introduced to various techniques and intricacies of Natural Language Generation.
4. Evaluation of the techniques can be done with automated metrics (Bleu, Rouge, Meteor).

Project Number: 5
Project Title: Paraphrase Identification & Generation
Mentor(s): Sagar (email: sagar.joshi@research.iiit.ac.in)
Details:

1. The project will constitute two tasks corresponding to two phases of the project timeline:

    a. Paraphrase identification - Experiment with several classification systems to determine if from a pair of sentences, either is a paraphrase of the other
    b. Paraphrase generation - Using the datasets used in phase 1, experiment with one or two sequence-to-sequence models to generate paraphrases
2. A multitude of datasets and metrics are available for students to choose from.
    a. Some popular datasets include PAWS, MSRP, QQP, PAN, etc.
    b. For paraphrase identification, usual classification metrics can be used.
    c. Metrics like BLEU, TER/TERp, METEOR, etc. can be used for estimating the quality of generated paraphrases.
3. Overall, the project will give an exposure to try hands on a range of libraries, models and metrics used in NLP tasks.

Project Number: 6
Project Title: Classification of COVID19 tweets containing symptoms
Mentor(s): Sravani & Tathagatha (email: sravani.boinepelli@research.iiit.ac.in & tathagata.raha@research.iiit.ac.in)
Details:

- Identifying personal mentions of COVID19 symptoms requires distinguishing personal mentions from other mentions such as symptoms reported by others and references to news articles or other sources. The classification of medical symptoms from COVID-19 Twitter posts presents two key issues: First, there is plenty of discourse around news and scientific articles that describe medical symptoms. While this discourse is not related to any user in particular, it enhances the difficulty of identifying valuable user-reported information. Second, many users describe symptoms that other people experience, instead of their own, as they are usually caregivers or relatives of people presenting the symptoms. This makes the task of separating what the user is self-reporting particularly tricky, as the discourse is not only around personal experiences.
- This task is considered a three-way classification task where the target classes are:
    - self-reports,
    - non-personal reports, and
    - literature/news mentions.
- Dataset: From Health Language Processing Lab @ Penn IBI, SMM4H 2021

Project Number: 7

Project Title: Multi-Label Classification of Mental Health Disorders
Mentor(s): Sravani & Tathagatha (email: sravani.boinepelli@research.iiit.ac.in & tathagata.raha@research.iiit.ac.in)
Details:

- Mental illnesses rank as some of the most disabling conditions, affecting millions of people, across the globe. In general, the main challenge of mental disorders is that they remain difficult to detect on suffering patients. In an online environment, the challenge extends to the implementation of proper algorithms to assist in the detection of such illnesses. Other challenges included with such multi label classification problems is the class imbalance of data. Investigate the impact of using deep learning architectures and emotional patterns identified by the clinical practitioners and computational linguists to enhance the prediction capabilities of mental illness detection models.
- Dataset(s):
    - SMHD contains nine mental health conditions including Depression, ADHD, Anxiety, Bipolar, PTSD, Autism, OCD, Schizophrenia, Eating with diagnosed users
    - BeyondBlue dataset (smaller dataset which includes depression, anxiety, PTSD & Trauma, Suicidal thoughts and self-harm etc)
    - Reddit Mental Health Dataset: 15 specific mental health support groups, 2 broad mental health subreddits, 11 non-mental health subreddits

---

Project Number: 8
Project Title: Multi-lingual sexism detection and classification
Mentor(s): Harika (email: harika.a@research.iiit.ac.in)
Details:

1. Need to classify “tweets” and “gab post” (in English and Spanish) according to the following two tasks**:**
    a. The first task is a binary classification. The students have to detect whether or not a given text (tweet or gab) is sexist (i.e., it is sexist itself, describes a sexist situation or criticizes a sexist behaviour).
    b. Once a message has been classified as sexist, the second task aims to categorize the message according to the type of sexism (according to the categorization proposed by experts and that takes into account the different facets of women that are undermined) into five categories such as Ideological and inequality, Stereotyping and dominance, Objectification, Sexual violence, Misogyny and non-sexual violence.
2. Source: EXIST task 2021

3. Evaluation will be done using standard metrics like Accuracy, F-macro, F-micro.

Project Number: 9
Project Title: Cross-lingual data-to-text for Indian languages
Mentor(s): Shivprasad Sagare (email: shivprasad.sagare@research.iiit.ac.in
)
Details:

1. Data-to-text generation is the task of generating text from a structured data source. Several datasets proposed in the recent past, and performance of the transformer based models have led to rise in popularity of this task. Although, as observed in several NLP tasks, such parallel data is not available in Indian languages. This project aims at performing following two subtasks:
    a. Implementing a data-to-text NLG system on webNLG dataset. webNLG is one of the standard datasets containing facts and sentences mapped together. It is available in English language.
    b. Exploring the ways to apply the data as well as models, on Indian languages. This can include translating the webNLG dataset in any Indian language e.g. Hindi and then building the models on it. Another approach can be to train a multilingual NLG model on the English webNLG dataset and then evaluate it on Hindi test data in zero-shot way.
2. Evaluation will be done using standard NLG metrics like BLEU, ROUGE applied on webNLG test data as well our inhouse test data for Indian languages. A qualitative analysis of the usability of the developed approaches is also expected.
3. This project gives the opportunity to work on increasingly popular NLP task of data-to-text, with well defined dataset, evaluation metrics, and resources available. It also has a research scope to it and will require working in the areas of multilinguality, cross-lingual zero-shot transfer, transformer based models, etc.
4. Links
    a. webNLG dataset WebNLG Challenge 2020 - WebNLG Challenges (loria.fr)
    b. Resources
        i. Example of a data-to-text system using pretrained language models https://www.aclweb.org/anthology/2020.inlg-1.14
        ii. Multilingual models, cross-lingual transfer, translate-train strategies A deep dive into multilingual NLP models - Peltarion

---

Project Number: 10
Project Title: Template generation from data-to-text
Mentor(s): Tushar (email: tushar.abhishek@research.iiit.ac.in)

- **Description**: Neural encoder-decoder models have significant success in data-to-text generation, but they often are uninterpretable and prone to hallucination. Instead of learning to directly generate text from structure data (tables, RDF, databases, etc.), the task is automatic generation of templates from a given <data, text> pair. This project includes evaluating and enhancing existing models (neural, statistical, rule-based, etc.) that can be adapted to different domains.
- **Dataset(s)**: Team is free to explore any data-to-text dataset (like WikiBIO, Rotowire, WeatherGov, E2E, WebNLG, ToTTO). Although,the E2E dataset is more appropriate for the task as it includes templatic sentences in the dataset.
- **Baselines**: Following baselines that can be referred for this task:
  - VARIATIONAL TEMPLATE MACHINE FOR DATA-TO TEXT GENERATION
  - Learning Neural Templates for Text Generation
- **Evaluation**: Although we can't evaluate the quality of the template as no ground truth is present for the task. Still, we can compare the ground truth text (in <data, text> pair) with the candidate text generated using the learned templates and structure data. For this you are free to choose any of standard text evaluation metrics like bleu, nist, rouge, meteor, bleurt, CIDEr.
- **Analysis**: Teams also need to do qualitative analysis on the learned templates.
- **Bonus Task (optional)**: An additional task can be performed of creating the above architecture for any one of the Indian languages that a team is familiar with.

Project Number: 11
Project Title: Political Ideology prediction - Identifying bias in news media
Mentor(s): Vijay (email: vijayasaradhi.i@research.iiit.ac.in)
Description: Identifying political bias in news media is very important. It can assist in identifying fake news
Dataset: Mediabiasfactcheck, allsides, nelagt etc any other which the team can create/source/find
Evaluation: can be done on F1 score as metric or any other suitable score depending on the kind of problem formulation

Analysis: This project has a very huge scope of analysing the data. Given that there are a lot of news sources and lot of content being produced(CCNEWS), it will give students a good chance of working things on bigger scale

---

Project Number: 12
Project Title: Clickbait intensity prediction and its relevance to Fake News
Mentor(s): Vijay (email: vijayasaradhi.i@research.iiit.ac.in)
Description: How can we measure the intensity of clickbaitiness of a news title/article. What makes it clickbaity and how can it assist in identifying fake news.
Dataset: Webis Clickbait Corpus 2017, NELAGT2019, NELAGT2020, FakeNewsNet etc.
Evaluation: Can be done with MSE for intensity prediction and F1 score for fake news prediction
Analysis: This project connects the clickbait intensity with the fake news prediction, a lot of scope for innovation like jointly learning the classifiers, adversarial settings etc.

Project Number: 13
Project Title: Author style profiling
Mentor(s): Vijay (email: vijayasaradhi.i@research.iiit.ac.in)
Description: How can we identify the author of a given text. The text can be an email, a chat message, or a news article.
Dataset: Pan Author profiling dataset, Enron email dataset, News dataset(to be crawled)
Evaluation: Can be done accuracy/F1 score depending on the type of formulation
Analysis: This project aims to identify the author of a given piece of text and predict the possible author.
Analysis: There is a huge scope of analysis as we have multiple datasets and new datasets can easily be sourced. This also has cyber-forensic applications also.

Project Number: 14
Project Title: Identifying the purpose of a citation in scientific literature
Mentor(s): Himanshu & Bhavyajeet (email: himanshu.maheshwari@research.iiit.ac.in & bhavyajeet.singh@research.iiit.ac.in)

Description:

1. The aim of this project is to identify the purpose of a citation in the scientific literature. The purpose could be either of these six: Background, Uses, Compare and Contrast, Motivation, Extension, and Future.
2. Thus this is a multiclass classification of citation purposes.
3. The dataset is very skewed and thus the existing models have a F1 score of only 0.26973 which is very less for practical purposes. Thus the aim of this project is to address this skewness and try to achieve better results.

Resources:
https://sdproc.org/2021/sharedtasks.html#3c
https://www.kaggle.com/c/3c-shared-task-purpose-v2/overview
https://arxiv.org/pdf/1904.01608.pdf

Project Number: 15
Project Title: Identifying the importance of a citation in scientific literature
Mentor(s): Himanshu & Bhavyajeet (email: himanshu.maheshwari@research.iiit.ac.in & bhavyajeet.singh@research.iiit.ac.in)
Description:

1. The aim of this project is to identify the importance of a citation in the scientific literature. There are two classes of importance: Incidental and Influential.
2. An incidental citation means that the citation is not very important in the flow of the paper and is just there to create an argument or for completion.
3. Influential citation means that the citation is important for the paper and we could not do without it.
4. This is a binary classification of citation importance.

Resources:
https://sdproc.org/2021/sharedtasks.html#3c
https://www.kaggle.com/c/3c-shared-task-purpose-v2/overview
https://arxiv.org/pdf/1904.01608.pdf

Project Number: 16
Project Title: Creating a parallel corpus for Layperson Summarization
Mentor: Sayar (sayar.ghosh@research.iiit.ac.in)
Description: Creation of a labelled dataset mapping scientific research papers to their corresponding lay summaries.

There are various online sources including blog posts, videos and podcasts describing technical papers in layperson terms restricting overly specific technical jargon. The task would be to identify such sources, crawl text in case of blog posts and collect transcripts in case of videos and podcasts. A mapping would be drawn from a paper to its corresponding scraped raw summary. The raw summaries would require some cleaning, transformation (might involve guidance from paper sections such as the 'abstract', 'introduction', and 'conclusion'), refinement, and post-processing (considering the evaluation and analysis outcomes) to serve as an approximate gold standard.
Dataset: Sample data: https://github.com/WING-NUS/scisumm-corpus/blob/master/README_Laysumm.md#sample-dataset
Evaluation and Analysis: (1) A statistical study of the linguistic features, readability, coherence, word usage, (2) A classifier to check whether a text piece can indeed serve as a Lay Summary, (3) Some manual evaluation.
Selected Resources:

1. 'TalkSumm': https://aclanthology.org/P19-1204/
2. LaySumm shared task: https://ornlcda.github.io/SDProc/sharedtasks.html
3. Parsing Research papers: https://github.com/allenai/science-parse
4. 'Summaformers' ~ work on generating extended and lay summaries: https://aclanthology.org/2020.sdp-1.39/

Project Number: 17
Project Title: Acronym Extraction and Disambiguation in Scientific Documents
Mentor(s): Tathagata and Sravani (email: sravani.boinepelli@research.iiit.ac.in & tathagata.raha@research.iiit.ac.in)
Description:
This project is running shared task at Scientific Document Understanding Workshop at AAAI'22. The project contains two subtasks:
1. To identify acronyms and their meanings (i.e.,long-forms) from the documents. For instance:

Input: Existing methods for learning with noisy labels (LNL) primarily take a loss correction approach.
Acronym: LNL
Long form: learning with noisy labels
2. to find the correct meaning of an ambiguous acronym in a given sentence. The input to the system is a sentence containing an ambiguous acronym. The systems are

expected to find the correct expanded form of the acronym given the possible expansions for the acronym. For instance:

Input Sentence: All systems use their IP address to introduce themselves to the network.
Input Candidate Long-forms: 1. Internet Protocol, 2. Intellectual Property
Output: Internet Protocol

Check this website for more details:
https://sites.google.com/view/sdu-aaai22/shared-task

Project Number: 18
Project Title: Intended Sarcasm Detection
Mentor(s): Tathagata and Sravani (email: sravani.boinepelli@research.iiit.ac.in & tathagata.raha@research.iiit.ac.in)
Description: Intended Sarcasm Detection
This project is a running shared task at SemEval-2022. This project contains three subtasks:
1. Given a text, determine whetherit is sarcastic or non-sarcastic(Course-grain classification)
2. Given a text, determine which ironic speech category it belongs to, if any. Ironic categories include:
- sarcasm: tweets that contradict the state of affairs and are critical towards an addressee;
- irony: tweets that contradict the state of affairs but are not obviously critical towards an addressee;
- satire: tweets that appear to support an addressee, but contain underlying disagreement and mocking;
- understatement: tweets that undermine the importance of the state of affairs they refer to;
- overstatement: tweets that describe the state of affairs in obviously exaggerated terms;
- rhetorical question: tweets that include a question whose invited inference (implicature) is obviously contradicting the state of affairs.
3. For every sarcastic sentence, an explanation is provided why it is sarcastic and a rephrase that conveys the same meaning non-sarcastically. For this subtask, given a sarcastic text and its non-sarcastic rephrase, i.e. two texts that convey the same meaning, determine which is the sarcastic one.

Dataset has been provided for English and Arabic however using the English dataset for this project should be enough.

Check this website for more details:
https://sites.google.com/view/semeval2022-isarcasmeval