Data Science Project

1. **SUPRIYO MANDAL (supriyo.pcs17@iitp.ac.in) Room No- 516(CSE Dept.), Block-3.**

   Project Title : Social Promoter Score based User-Item Interaction: A New Perspective in Recommendation.

   Project Description :
   Most of the existing recommender systems understand preference degree of users based on user-item interaction ratings with implicit feedback. The rating based recommendation systems always ignore the negative users (who give bad ratings) who are also reliable reviewers. There are two types of negative users. Some negative users give negative or bad ratings randomly and some negative users who are reliable also, give ratings according to the quality of items. Similar characteristics are also applicable to positive users. From bad reflection of a user to a specific item, existing recommender systems presume that this item is not the user's preferred category. The bad reflection of a user to a specific item cannot be interpreted as a wrong category recommendation. **We have to investigate whether the item is not her preferred category or she is dissatisfied with the quality of her preferred item or she gives rating randomly and this is the main part of our work**.

   Model: SVD or PMF or MLP ---- (based on students' preferences)

   Useful Resource :
   Amazon dataset : http://jmcauley.ucsd.edu/data/amazon/
   Before MidSem:

   - Details study on Dataset
   - Understand review network
   - Calculate users' rating value, reliability score and SPS score

   After MidSem:

   - Develop Model and Apply Matrix Factorization
   - Predict rating score based on users' characteristic
   - Analysis performances.

2. Rimjhim(rimjhim.pcs16@iitp.ac.in), Ph. No- 8434444196, Room No-516(CSE Dept.), Block-3.

A. Data Wrangling

BEFORE MIDSEM

Foursquare is LBSN whose data has been used extensively in countless researches. Unfortunately, data of Foursquare is not freely available. However, there is a hack of collecting Foursquare data using Twitter data as users also share their Foursquare checkings on Twitter. The task of this project if to collect Foursquare checkin data via Twitter platform. The checkin data must contain information like, checkin time, checkin user, checkin place(lat,long). For many analyses these data needs to be enriched with other metadata such as place name, place popularity, etc. Foursquare provide details of various places which can be collected via Foursquare API.

AFTER MIDSEM

Second task of this project is to enrich checkin data with place information, in case place information is not available for a place, keep a tag of 'NotAvail'. For other sample data and other queries, contact me.

B. Clustering: Unsupervised Learning

BEFORE MIDSEM

Data Science often required grouping of data on basis of few parameters such as similarity, closeness, etc. Here is a very similar task of clustering words into 9 clusters. The name of the cluster are Arts & Entertainment, College & University, Food, professional & other Places, Nightlife Spot, Recreation & Outdoors, Shop & Service, Travel & Transport, Residence. This can be done on the basis of semantic similarity of a word to the cluster. For example the word coffee, is more similar to the cluster food than any other cluster. For finding the semantic similarity, use cosine similarity between word vectors. For obtaining word vectors use Google pre-trained News Vectors(https://github.com/mmihaltz/word2vec-GoogleNews-vectors).

AFTER MIDSEM

You can take help of other dictionaries for the same such as nltk dictionary and category dictionary. This step needs is for doing the same task from other resources as well. For further details, contact me.

C. Popularity Analysis of different POI(Point-of-Interest) with time.

BEFORE MIDSEM

Day by day many POI are coming and many are even going. Similarly the popularity of POI also keeps on changing. Given a global dataset of Foursquare, indentify the places which have a sharp change in the popularity during a period of 22 months. The popularity is defined in terms of number of checkins at a particular POI, if the number of checkins decreases or increases than a median checkin rate, there is a popularity change.

AFTER MIDSEM

Identify POI with overall very high popularity, overall very low popularity, significant popularity change POIs. Compare and contrast the popularity change among two asian countries vs two Western Countries. For further details, contact me.

3. Manish Bhanu (manish.pcs16@iitp.ac.in, 6204696501)

Proposed work: **Prediction of Traffic volumes in a city.**

An Intelligent Transportation System forms an integral component of a smart city. Formulation of traffic policy and well regulated traffic operation facilitates smooth mobility behaviour in the city. Traffic volume prediction is a challenging research field when the whole city is concerned. Our goal is to best approximate the prediction at multiple source and destination pairs of the city and for few time steps ahead in the future. We look at the best utilisation of the available resources to improve the prediction. For our objective, we would look into some feature engineering techniques like dimensionality reduction, data-imputation and data-analysis concept like classification, clustering etc and few matrix based concepts SVD, PCA etc. Few forecasting concepts using ANN, ARIMA etc.

Mid-sem Target: Apply SVD and PCA on the features extracted from the traffic data.
End-sem Target: Apply ew forecasting concepts using ANN, ARIMA etc.

4. Anita Chandra, email id (email id: anita.pcs15@iitp.ac.in), phone No:-8292805149

1. Identifying most influential videos in an educational Youtube channel like NPTEL.

Mid Sem Target : Preparation of data -

Crawl the Youtube dataset using YouTube API. Write a python code to extract details of a given channel id. Channel should be educational channel, for example, coursera, Yale, MIT OpenCourseWare, NPTEL, Khan academy etc. For more educational channels follow this https://www.shoutmeloud.com/best-youtube-channels-learning-development-growth.html.

1. Crawl the YouTube data using Youtube API. Give a channel id, find out all videos_id, rate, comments, users details. (I have code, I will provide it)
2. Apply ML algorithms to identify most influential videos of a particular channel. Please refer https://www.sciencedirect.com/science/article/abs/pii/S0360131518302392 to understand the importance of analysing educational channels.

2. Application of summarization of graph.

Mid Sem Target: Apply already existing proposed graph summarization algorithms:

Apply different existing graph summarization techniques or algorithms (eg., Sampling by random node selection, Sampling by random edge selection, Sampling by exploration) to get summarized graph. You can read basics, requirements of graph summarisation and different summarisation algorithms from here https://www.slideshare.net/aftabalam18/a-graph-summarization-a-survey-summarizing-and-understanding-large-graphs

https://slideplayer.com/slide/12389646/

https://cs.stanford.edu/people/jure/pubs/sampling-kdd06.pdf

Then, understand the algorithms and apply on freely available online datasets. Yo will get lots of datasets from here http://konect.uni-koblenz.de/.

1. Study and apply several existing summarization algorithms for graph. To understand the concepts and applications of graph summarization, please go through the paper given in this link: https://tsafavi.github.io/assets/pdf/liu-2018-graph-summarization.pdf
2. Apply summarized graph to check does it preserve the original properties of graph.

5. Saswata Roy (saswata.pcs17@iitp.ac.in/rhonson@gmail.com) (7761807224)

A. First Project

Project Title: ***Detection of Fake News on Social Media through propagation path classification using Convolutional neural networks (CNN)***

Project Description: We are given a set of news stories (here source tweets) and for each of these news, we have to predict whether this news shall become fake or not in future. We have two twitter datasets (a) Twitter15 and (b) Twitter16 dataset. Each dataset contains four different labels, i.e., “fake”, “true”, “unverified”, and “debunking of fake”. You can only use user information corresponding to each news. You have to build a Convolutional Neural Network based classification model which will use these user information as input. Goal of this classification model is to predict the label of these news.

The concept of “**propagation path**” is as follows. Let A = $\{a_1, a_2, . . . , a_{|A|}\}$ be a set of news stories, U = $\{u_1, u_2, . . . , u_{|U|}\}$ be a set of social media users. Each user $u_j \in U$ is associated with a user vector $x_j \in R^d$, which represents the characteristics of the user. We define the propagation path of a given news story a_i as a variable-length multivariate time series $P(a_i) = . . . , (x_j, t), . . . ,$ in which each tuple (x_j, t) denotes that user u_j tweets/retweets the news story a_i at time t. In this project, we set the time of a source tweet being posted to 0. Thus, $t > 0$ refers to the time of a retweet being posted.

Models Used: Recurrent Neural Network.

Useful Resource: “Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks ”

Dataset: https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0

Before midsem: Implement this task using simple machine learning algorithm such as random forest and Support vector machine

After Midsem: Use CNN

B. Second Project

Project Title: ***Detection of Fake News on Social Media leveraging propagation path classification and textual content using RNN***

Project Description: We are given a set of news stories (here source tweets) and for each of these news, we have to predict whether this news shall become fake or not in future. We have two twitter datasets (a) Twitter15 and (b) Twitter16 dataset. Each dataset contains four different labels, i.e., “fake”, “true”, “unverified”, and “debunking of fake”. You have to use user information corresponding to each tweet as well as tweet content. You have to build two Recurrent Neural Network (RNN) based classification models. One RNN model takes user information and another takes textual content as input. Output of this two RNN model will be concatenated. Goal of this classification model is to predict the label of these news.

The concept of this propagation path is as follows. Let A = $\{a_1, a_2, . . . , a_{|A|}\}$ be a set of news stories, U = $\{u_1, u_2, . . . , u_{|U|}\}$ be a set of social media users. Each user $u_j \in U$ is associated with a user vector $x_j \in R^d$, which represents the characteristics of the user. We define the propagation path of a given news story a_i as a variable-length multivariate time series $P(a_i)$ = . . . , (x_j, t), . . . , in which each tuple (xj, t) denotes that user u_j tweets/retweets the news story a_i at time t. In this project, we set the time of a source tweet being posted to 0. Thus, $t > 0$ refers to the time of a retweet being posted.

Models Used: Recurrent Neural Network.

Useful Resource: “Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks ”

Dataset: https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0

Before midsem: Implement this task using simple machine learning algorithm such as Logistic Regression and KNN

After Midsem: Use RNN

C. Third Project

Project Title: ***Rumor Detection on Twitter with Tree-structured Recursive Neural Networks***

Project Description: We define a Twitter rumor detection dataset as a set of claims C = $\{C_1, C_2, \cdot\cdot\cdot, C_{|C|}\}$, where each claim C_i corresponds to a source tweet r_i which consists of ideally all its relevant responsive tweets in chronological order, i.e., $C_i = \{r_i, x_{i1}, x_{i2}, \cdot\cdot\cdot, x_{im}\}$ where each x_{i*} is a responsive tweet of the root r_i. Note that although the tweets are notated sequentially, there are connections among them based on their reply or repost relationships, which can form a propagation tree structure with r_i being the root node. We have to formulate this task as a supervised classification problem, which learns a classifier f from labeled claims, that is f : C_i ->

Y_i, where Y_i takes one of the four finer-grained classes: non-rumor, false rumor, true rumor, and unverified rumor. We have two twitter datasets (a) Twitter15 and (b) Twitter16 dataset.

Models Used: Recursive neural models based on a bottom-up and a top-down tree-structured neural networks

Useful Resource: "Rumor Detection on Twitter with Tree-structured Recursive Neural Networks "

Dataset: https://www.dropbox.com/s/7ewzdrbelpmrnxu/rumdetect2017.zip?dl=0

Before midsem: Implement this task using simple machine learning algorithm Decision Tree

After Midsem: Use ***Tree-structured Recursive Neural Networks***

D. Fourth Project

Project Title: ***Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter***

Project Description: We consider a collection D of rumours, $D = \{R_1, \cdots, R_{|D|}\}$. Each rumour R_i contains a set of tweets discussing it, $R_i = \{d_1, \cdots, d_{ni}\}$. Each tweet is represented as a tuple $d_j = (t_j, W_j, m_j, y_j)$, which includes the following information: t_j is the posting time of the tweet, W_j is the text message, m_j is the rumour category and y_j is the label, $y_j \in Y = \{\text{supporting, denying, questioning, commenting}\}$. We have to define the stance classification task such that in which each tweet d_j needs to be classified into one of the four categories, $y_j \in Y$, which represents the stance of the tweet d_j with respect to the rumour R_i it belongs to.

Models Used: *Hawkes Processes*

Useful Resource: "Hawkes Processes for Continuous Time Sequence Classification: an Application to Rumour Stance Classification in Twitter "

Dataset: provided in the paper.

Before midsem: Implement this task using simple machine learning algorithm SVC with RBF kernel

After Midsem: Use ***Hawkes Processes***

E. Fifth Project

Project Title: ***Detect Rumor and Stance Jointly by Neural Multi-task Learning (MTL)***

Project Description: Our goal is to formulate a multi-task model that jointly learns the rumor detection and stance classification models, where one task may or may not use data from the same source as the other. For instance, we can typically use tweets in rumor detection but use news reports in stance classification, considering the availability of training data and specific setting. Since tweets are short in nature, containing very limited context, a claim is generally associated with a collection of posts that are relevant to it. Therefore, we model the Twitter data as a set of claims $\{C_1,C_2, \cdot\cdot\cdot ,C_{|C\,|} \}$, where each claim $C_i = \{(x_{ij},t_{ij})\}$ is composed of a set of relevant tweets, and x_{ij} is a post posted at time t_{ij}

Models Used: RNN-based neural multi-task model

Useful Resource: “Detect Rumor and Stance Jointly by Neural Multi-task Learning ”

Dataset: Pheme and Rumor Eval 17

Before midsem: Detect rumor and stance separately.

After Midsem: Use MTL to Detect Rumor and Stance Jointly

G. Sixth Project

Project Title: Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure

Project Description: We defined the rumour stance detection problem as a simple four-way classification task, where every tweet in the dataset (source and direct or nested reply) should be classified into one among four classes: support, deny, query, and comment.

Features Used: (1) **structural features** (Retweet Count: The number of retweet of each tweet. Question Mark: presence of question mark "?"; binary value and so on.) (2) **Conversation Based Features** (Text Similarity to Source Tweet: Jaccard Similarity of each tweet with its source tweet)**((3) Affective Based Features (4) Dialogue-Act Features**

Models Used: Support vector classifier with radial basis function (RBF) kernel, Naive Bayes, Decision Trees, Support Vector Machine.

Useful Resource: “Stance Classification for Rumour Analysis in Twitter: Exploiting Affective Information and Conversation Structure ”

Dataset: provided in the paper.

Before midsem: Finding out all these Affective Information and Conversation Structure feature

After Midsem: Implement classification algorithms

H. Seventh Project

Project Title: Online clustering and classification for real-time event detection in Twitter

Project Description: Initially we have to continuously keep on collecting Twitter data using a set of violence/riot related keywords (this keyword set will be given to you). Next, we apply to each message common pre-processing functions in order to remove irrelevant contents. Next, we run an online clustering algorithm that groups similar messages together into a set of clusters C, where each cluster c_j is represented by its centroid r_j. At the end of each slot, we analyze C with a binary classifier in order to detect whether it contains an event or not. For the scope of this paper we assume that each slot contains at most one event, but the methodology can be easily extended to multiple events per slot. If the classifier detects an event, we assign a score to each cluster cj and we select the highest in rank to identify the cluster that contains the event. Finally, we compare its content with a set of user-defined topics in order to decide whether the event (cluster) matches or not.

Models Used: SVM classifiers, incremental online clustering algorithms/k-means clustering algorithms

Useful Resource: "Online clustering and classification for real-time event detection in Twitter "

Before midsem: Do online clustering

After Midsem: Do classification and rest of the work.

H. Eighth Project

Project Title: Summarization and Sentiment analysis of Violence/ Radicalism related events in Twitter data

Project Description: From a given twitter corpus, you have to clean the data first. After preprocessing, you are supposed to run k-means clustering algorithms over these dataset. Out of these clusters, you have to select only top k clusters. Then you have to show the percentage of positive, negative and neutral sentiments in each cluster. You have to prepare wordcloud of each cluster also (where *a wordcloud is a visualization wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes).*

Models Used: k-menas clustering algorithms using scikit learn python

Useful Resource: https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/

Dataset: Will be provided

Before midsem: Apply different clustering algorithms

After Midsem: Do remaining work

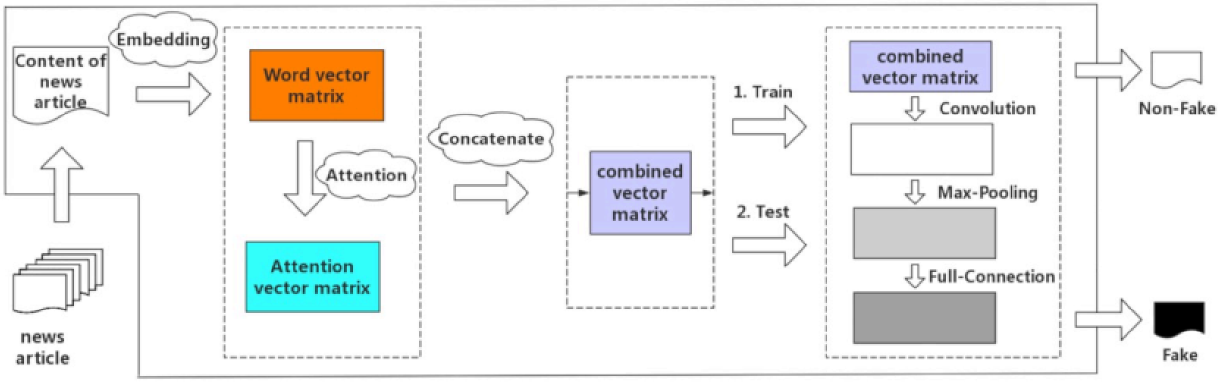
I. Ninth Project

Project Title: Self Multi-Head Attention-based Convolutional Neural Networks for fake news detection

Project Description: Suppose you are given a set of news articles. You have to build first a word vector model replaces each word in a news with its corresponding vector and creates a news article matrix X $\in R^{n \times d}$ where n is the number of words in the article, and the d is the dimension of word embedding,

$$X = (x_1, x_2, \ldots, x_i)$$

where $x_i \in R^d$ is the d-dimensional word vector corresponding to the i-th word in the news. Then you have to leverage self multi-head attention mechanism to obtain the internal spatial relationship in words. Then word vector matrix and attention matrix are concatenated to produce combined vector matrix which is then fed to convolution network for predicting the label of news article (Label is either fake or non-fake).



Models Used:

Useful Resource: "Self Multi-Head Attention-based Convolutional Neural Networks for fake news detection"

Dataset: https://www.kaggle.com/mrisdal/fake-news#fake.csv

Before midsem: Create your own word2vec model

After Midsem: apply mentioned deep learning model.

Mentor: Shalini(shalini.pcs16@iitp.ac.in)

Note: Datasets and other information will be made available after the project assignment

1. **Project title:** Using Context Information for Dialog Act Classification in DNNFramework

Description: Dialog act (DA) represents a function of a speaker's utterance in either human-to-human or human-to-computer conversations. Correct identification of DAs is important for understanding human conversations, as well as for developing intelligent human-to-computer dialog systems (either written or spoken dialogs). For example, recognizing DAs can help identify questions and answers in meetings, customer service, online forum, etc. Intuitively we would expect that leveraging dia-log context can help classify the current utterance. Propose a technique that incorporate context information for DA classification over the baseline method of using convolutional neural networks (CNN) for sentence classification.

Mid-term Target: (a) a hierarchical RNN/LSTM and CNN to model the utterance sequence in the conversation, where the input to the higher level LSTM and CNN unit is the sen-tence vector from the sentence level CNN model.

End-sem Target:

(a) a two-step approach where the predicted DA results for the previous utterances, either labels or probability distributions, are concatenated with the sentence CNN vector for the current utterance as the new input for classification; (b) sequence level decoding based on the predicted DA prob-abilities and the transition probabilities betweenDA labels. Some of these methods have not been exploited previously for this task.

2. **Project title:** A Machine Learning Model for Stock Market Prediction

Description: Stock market prediction is the act of trying to determine the future value of a company stock or other financial instrument traded on a financial exchange. The successful prediction of a stock's future price will maximize investor's gains. Propose a machine learning model to predict stock market price. The proposed algorithm should integrates Particle swarm optimization (PSO) and least square support vector machine (LS-SVM). Employ PSO algorithm to optimize LS-SVM to predict the daily stock prices. Proposed model should be based on the study of stocks historical data and technical indicators. Use PSO algorithm selects best free parameters combination for LS-SVM to avoid overfitting and local minima problems and improve prediction accuracy.

Mid-term Target: Apply SVM and LSTM models for stock market prediction.
End-term Target: Use PSO to improve the prediction accuracy.

3. **Project title:** Image classification: Zero-shot Image Recognition Using RelationalMatching, Adaptation and Calibration
Project description: Zero-shot learning (ZSL) for image classification focuses on recognizing novel categories that have no labeled data available for training. The learning is generally carried out with the help of mid-level semantic descriptors associated with each class. This semantic-descriptor space is generally shared by both seen and unseen categories. However, ZSL suffers from hubness, domain discrepancy and biased-ness towards seen classes. To tackle these problems, propose a three-step approach to zero-shot learning. Firstly, a mapping is learned from the semantic-descriptor space to the image-feature space. This mapping should learns to minimize both one-to-one and pairwise distances between semantic embeddings and the image features of the corresponding classes. Secondly, propose a test-time domain adaptation to adapt the semantic embedding of the unseen classes to the test data. This will be achieved by finding correspondences between the semantic descriptors and the image features. Thirdly, propose scaled calibration on the classification scores of the seen classes. This is necessary because the ZSL model is biased towards seen classes as the unseen classes are not used in the training.

Mid-term Target: Firstly, a mapping is learned from the semantic-descriptor space to the image-feature space. This mapping should learns to minimize both one-to-one and pairwise distances between semantic embeddings and the image features of the corresponding classes.
End-term Target: Secondly, propose a test-time domain adaptation to adapt the semantic embedding of the unseen classes to the test data. This will be achieved by finding correspondences between the semantic descriptors and the image features. Thirdly, propose scaled calibration on the classification scores of the seen classes.

4. **Project title:** Creating story chains using Reddit comments
Description: Large amounts of information about news events are published on the Internet every day in online newspapers. Compose a story chain from the set of reddit comment that coherently connect together to help answer the question of how event A is related to event B. Utilize the fact that news events are composed of “Who”, “What”, “Where” and “When” for finding the news story. Prepare the gold standard for a sample of dataset (50 news articles) and check the correctness of your technique.

Mid-term Target: Find who, what, where and when in the news article.
End-term Target: Create the final story chain.

5. Project title: Classifying and Summarizing Information from Microblogs During Epidemics
Project description: During a new disease outbreak, frustration and uncertainties among affected and vulnerable population increase. Affected communities look for known symptoms, prevention measures, and treatment strategies. On the other hand, health organizations try to get situational updates to assess the severity of the outbreak, known affected cases, and other details.Recent emergence of social media platforms such as Twitter provide convenient ways and fast access to disseminate and consume information to/from a wider audience. Research studies have shown potential of this online information to address information needs of concerned authorities during outbreaks, epidemics, and pandemics. In this work, we target three types of end-users (i) vulnerable population—people who are not yet affected and are looking for prevention related information (ii) affected population—people who are affected and looking for treatment related information, and (iii) health organizations—like WHO, who are interested in gaining situational awareness to make timely decisions. Use Twitter data from two recent outbreaks (Ebola and MERS) to build an automatic classification approach useful to categorize tweets into different disease related categories. Moreover, Use the classified messages to generate different kinds of summaries useful for affected and vulnerable communities as well as health organizations.

Mid-term Target: classifying information from Microblogs During Epidemics
End-term Target: Summarizing Information from Microblogs During Epidemics

6. Project title: Summarizing Situational Tweets in Crisis Scenarios:An Extractive-Abstractive Approach
Description: Microblogging platforms such as Twitter are widely used by eyewitnesses and affected people to post situational updates during mass convergence events such as natural and man-made disasters. These crisis-related messages disperse among multiple classes/categories such as infrastructure damage, shelter needs, information about missing, injured, and dead people. Moreover, it's observed that sometimes people post information about their missing relatives and friends with personal details such as names and last seen location. The information requirements of different stakeholders (government, NGOs, and rescue workers) also vary a lot. This brings two fold challenges: 1) extracting important high-level situational updates from these messages, assigning them appropriate categories, and finally summarizing big trove of information in each category and 2) extracting small-scale time-critical sparse updates related to missing or trapped people. Propose a classification-summarization framework which first assigns tweets into different situational classes and then summarizes those tweets. In the summarization

phase, propose a two-step extractive-abstractive summarization framework. In the first step, extracts a set of important tweets from the whole set of information, develops a bigram-based word-graph from those tweets, and generates paths by traversing the word-graph. Next, it uses an optimization technique based on integer linear programming (ILP) to select the most important tweets and paths based on different optimization parameters such as informativeness and coverage of content words. Apart from general class wise summarization, also show the customization of our summarization model to address time-critical sparse information needs (e.g., missing relatives).

Mid-sem Target: Summarizing Situational Tweets in Crisis Scenarios:An Extractive Approach
End-sem Target: Summarizing Situational Tweets in Crisis Scenarios:An Abstractive Approach

7. Project title : Summarizing Microblogs during Emergency Events: A Comparison of Extractive Summarization Algorithms
Project description: Microblogging sites, notably Twitter, have become important sources of real-time situational information during emergency events. Since hundreds to thousands of microblogs (tweets) are generally posted on Twitter during an emergency event, manually going through every tweet is not feasible. Hence, summarization of microblogs posted during emergency events has become an important problem in recent years. Several summarization algorithms have been proposed in literature, both for general document sum-marization, as well as specifically for summarization of microblogs. In this work, evaluate and compare the performance of extractive summarization algorithms in the application of summarizing mi-croblogs posted during emergency events.

Mid-sem Target: Summarizing Situational Tweets in Crisis Scenarios: two Extractive summarization Approach
End-sem Target: Summarizing Situational Tweets in Crisis Scenarios: Propose an extractive summarization Approach

8. Project title: A Graph-based Approach for Detecting Critical Infrastructure Disruptions on Social Media in Disasters
Description: The objective is to propose and test a graph-based approach for detection of critical infrastructure disruptions in social media data in disasters. Understanding the situation and disruptive events of critical infrastructure is essential to effective disaster response and recovery of communities. The potential of social media data for situation awareness during disasters has been highlighted in recent studies. However, the application of social sensing in detecting disruptions of critical infrastructure is limited because existing approaches cannot provide

complete and non-ambiguous situational information about critical infrastructure. Therefore, to address this methodological gap, develop a graph-based approach including data filtering, burst time-frame detection, content similarity and graph analysis.

Mid-sem Target: develop a graph-based approach including data filtering, burst time-frame detection,
End-sem Target: content similarity and graph analysis.

9. Project Title: Representing Text for Joint Embedding of Text and Knowledge Bases
Description: Representing information about real-world enti-ties and their relations in structured knowledge base (KB) form enables numerous applications. Models that learn to represent textual and knowledge base relations in the same con-tinuous latent space are able to perform joint inferences among the two kinds of relations and obtain high accuracy on knowl-edge base completion. Propose a model that captures the compositional structure of textual relations, and jointly optimizes entity, knowledge base, and textual relation representations.

Mid-sem Target: Propose a model that captures the compositional structure of textual relations, and jointly optimizes entity.
End-sem Target: knowledge base, and textual relation representations.

10. Project Title: A lightweight and multilingual framework for crisis information extraction from Twitter data
Description: Obtaining relevant timely information during crisis events is a challenging task that can be fundamental to handle the consequences deriving from both unexpected events (e.g., terrorist attacks) and partially predictable ones (i.e., natural disasters). Even though microblogging-based online social networks (e.g., Twitter) have become an attractive data source in these emergency situations, overcoming the information overload deriving from mass events is not trivial. The aim of this work is to enable unsupervised extraction of relevant information from Twitter data during a crisis event, offering a lightweight alternative to learning-based approaches. Propose a *lightweight crisis management framework* that will integrate natural language processing and clustering techniques in order to produce a ranking of tweets relevant to a crisis situation based on their informativeness.

Mid-sem Target: Use NLP and clustering approach the form the clusters of tweets.
End-sem Target: Rank the tweets based on the informativeness.