Information Retrieval
Term Project Report

Mentor: Rajdeep Mukherjee

Group 17

Jalend Bantupalli	18EC10023
Ganesh Shiridi Balaji Udayagiri	18EC35010
Chapala Chinnikrishna Naidu	18EC10010
Kandi Vishnu Vardhan Reddy	18CS30022

Common task

Problem Statement:

Incorporate *stance classification* from “All-in-one: Multi-task Learning for Rumour Verification” into Tree LSTM-based *rumour detection* from “Going Beyond Content Richness: Verified Information Aware Summarization of Crisis-Related Microblogs” and obtain the results on PHEME-RNR dataset

Implementation done so far:

Mid evaluation

The individual papers are run the datasets given. Verified Information Aware Summarization of Crisis-Related Microblogs on PHEME-RNR dataset and All-in-on: Multitasking for Rumour Verification on already preprocessed data.

The stance labels are added to the trees generated as a result of generate_trees.py

End evaluation

Stance classification is added to every non root node, along with rumour detection at root node.

Results:

Verified summarisation:

K	T	numItr
30	10	1000

IN_FEATURES	OUT_FEATURES	NUM_ITERATIONS	BATCH_SIZE	HIDDEN_UNITS	LEARNING_RATE
80	2	10	50	128	0.001

Stance Classification at every non root node: (40 FEATURES)

Eval/Data	charliehebdo	germanwings-crash	ottawashooting	sydneysiege
Accuracy	0.72018779342	0.664688427299	0.677146311970	0.684684684684
F1-score	0.245825736186	0.199643493761	0.201874549387	0.203208556149

Stance Classification at every non root node: (80 FEATURES and weights (1/sqrt(freq)) for each stance class

Eval/Data	charliehebdo	germanwings-crash	ottawashooting	sydneysiege
Accuracy	0.738967136150 2348	0.673590504451 0386	0.678355501813 7848	0.684684684684 6847
F1-score	0.255034935144 39207	0.201241134751 77308	0.202089337175 7925	0.203208556149 7326

Rumour classification

40 Features

Eval/Data	charliehebdo	germanwings-crash	ottawashooting	sydneysiege
Accuracy	0.7768304914744233	0.5161290322580645	0.4648711943793911	0.5735042735042735
F1-score	0.4522015222763634	0.514682723483094	0.31734612310151883	0.3644758283541553

80 features

Eval/Data	charliehebdo	germanwings-crash	ottawashooting	sydneysiege
Accuracy	0.7311935807422267	0.5583126550868487	0.6085470085470085	0.6085470085470085
F1-score	0.6258639423777038	0.5334677419354840	0.4786562941458042	0.4786562941458042

Individual task

Problem Statement:

Run the assigned paper “Cascade-LSTM: A Tree-Structured Neural Classifier for Detecting Misinformation Cascades” on PHEME-RNR dataset

Implementation done so far:

The code was run for the provided dataset. - FalseNews_Code_Data_

PHEME-RNR dataset is preprocessed to suit the needs.

The code flow was analysed to find fault with the preprocessed dataset

A masked Cascade LSTM is developed for masking some user defined features as mentioned in the table below.

Raw_data_anon.csv	PHEME-RNR
tid	id
veracity	True/False/Unverified
cascade_id	available
rumor_id	available
rumor_category	Depending on one of the events, manually should assign - {Politics, War/Terrorism/Shootings, Viral}
parent_tid	parent_tweetid
tweet_date	created_at
user_account_age	(mask)
user_verified	TRUE/FALSE (mask)
user_followers	Followers_count (mask)
user_followees	mask
user_engagement	mask

cascade_root_tid	available same as tweet root id
was_retweeted	At least one child then 1 else 0

emotions_anon.csv	**PHEME-RNR**
tweet_id	id(mask)
sadness	mask
anticipation	mask
disgust	mask
surprise	mask
anger	mask
joy	mask
fear	mask
trust	mask
misc	mask

- **DGL library** had outdated functions - so functions had to be updated and **brought to the latest version** accordingly
- Made necessary changes in the code **(Cascade LSTM)** to run it on the given dataset.
- Masked the user defined features and tested on the given dataset.

Results:

The training and testing on original data using masked cascade lstm

Started experiment 11_19_2021__18_52_43__470569
Model 11_19_2021__18_52_43__470569 saved with test AUC 0.5000 | train AUC 0.4703 at epoch 0
Experiment 11_19_2021__18_52_43__470569 terminated with test AUC 0.5000 at epoch 9

Experiments done

1. **PHEME-RNR doesn't have the required features** - so the extra features are computed if available and user defined features are removed in the cascades

- **Assumption that reactions are retweets - csv is created based on the assumptions made mentioned above**
- Due to this assumption, we found out there was a loop in the graph created -
- dgl._ffi.base.DGLError: [04:05:18] /tmp/dgl_src/src/array/cpu/./traversal.h:222: Error in topological traversal: loop detected in the given graph

2. It was figured out that cascade_root_id is not the cascade id of the root, rather it is root id of the tree is it a part of.

- **Assumption that reactions are retweets - csv is created based on the assumptions made mentioned above**
- Due to this assumption, we found out there was a loop in the graph created -
- dgl._ffi.base.DGLError: [04:05:18] /tmp/dgl_src/src/array/cpu/./traversal.h:222: Error in topological traversal: loop detected in the given graph

3. Was_retweeted was set to 1 to only those which have atleast one child.

Assumption that reactions are retweets - csv is created based on the assumptions made mentioned above

- Due to this assumption, we found out there was a loop in the graph created -

- dgl._ffi.base.DGLError: [04:05:18] /tmp/dgl_src/src/array/cpu/./traversal.h:222: Error in topological traversal: loop detected in the given graph

4. Removed the loops manually
- dgl._ffi.base.DGLError: [01:02:08] /tmp/dgl_src/src/array/cpu/spmat_op_impl_coo.cc:461: Check failed: thread_prefixsum[num_threads] == NNZ (24 vs. 26) :
- Tried to solve this issue but the resources available are very limited.