Rumour detection from Posts and Comments

Reference: PostCom2DR – Utilizing information from post and comments to detect rumours
by Yanjie Yang, Yuhang Wang, Li Wang, Jie Meng

Report created by:
Team Strawhats
Members:
Ritik Shrivastava (214101065)
Divyanshu Nauni (214101063)
Patel Miki Maheshbhai (214101036)

Phase 1:

1. Introduction

The development of social media has allowed to share information in fast and efficient way. The social media has also become prominent platform for publishing and spreading rumours. Most social media contains comments, these comments can be used to utilize to detect rumours. The reply structure, mutual selection formation between post and comments, topic drift within comments also helps with detecting rumour apart from comment and post content information. In the paper PostCom2Dr they have proposed a rumour detection model that uses the structure of comments along with comment content of comments and post. Paper has proposed the bilevel GCN and self-attention mechanism to learn the representation of comments. The post-comment co-attention mechanism is introduced to selectively fuse information, and this helps the model focus on more relevant information. Apart from this, CNN is built to capture the local topic drift on time series inside the comments. Then the global and local representation is concatenated to detect rumour. The experiment is conducted on FakeNewsNet and Twitter 15. The results of our experiments is shown below.

2. Challenges in research problem:

The research problem is about determining whether the news is rumour or not. Detecting rumour on social media can be done using various methods. These methods are divided into three categories. These categories are discussed below:

2.1.1 Content based methods:

This method focuses on the text features extracted by the post. The classification can be done using machine-learning methods using the features from the post. The focus is on language features of tweet such as special characters, emoticons, positivity or negativity of the words, tags, etc. This method requires people to manually design and extract features from the datasets. This makes it time consuming. The content can be from different domains too. So the features that work on one domain may not work on another.

2.1.2 User based methods:

This method studies the behaviour of user to identify the likeliness of posting rumour. Information like name, verification of account, activeness of account, etc are used to determine the credibility of the post.

2.1.2 Propagation based methods:

Rumour is written in a way that the writer tries to imitate the actual post as much as possible. This makes detecting rumour difficult. The comments on post creates a propagation network of post. Repost based methods are the propagation based methods that analyse the repost structure, writing style, etc. Comment based method is also another type of propagation based method that uses comments to analyse whether the post can be a rumour or not.

This can be done using various models. These models are discussed below:

2.2.1 DTC:

This is a rumour detection method which uses decision tree classifier on various features to obtain information credibility. The features can be Message based (length of message, punctuation information, emoticon or URL information, etc), User-based (age, followers, verified or not, etc), Propagation based (propagation max or avg degree, max or avg depth, etc). These features are the input to decision tree classifier for classification.

2.2.2 SVM – RBF:

This SVM based RBF kernel uses overall statistics of posts. Apart from the features mentioned in above part, it also uses Client-based features (device information), Location-based features. Then classification is done using stable and effective SVM method.

2.2.3 CNN:

This model creates 1-d convolution layer with a filter size 3 over the embeddings of the posts and comments, followed by max-pooling layer and a fully connected layer.

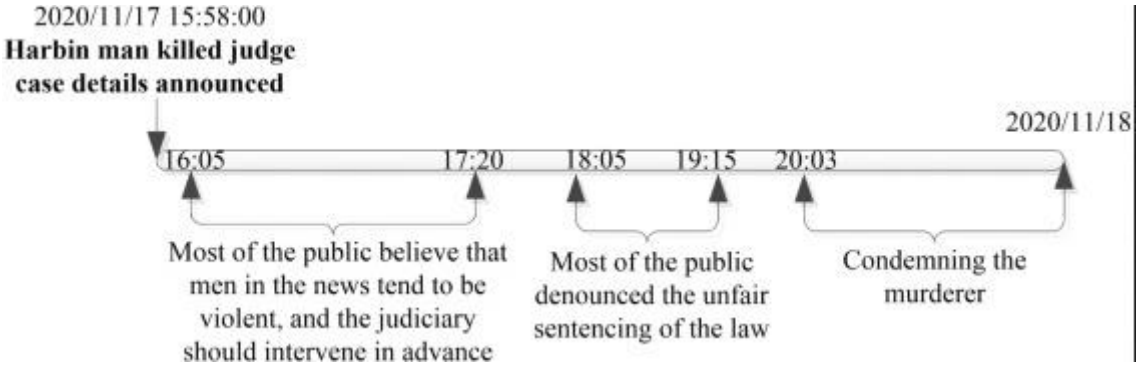
2.2.4 dEFEND:

This is the latest rumour detection model is based on comments and posts which can learn the correlation between text sentences and comments. It uses co-attention between comment and post for fake news detection.

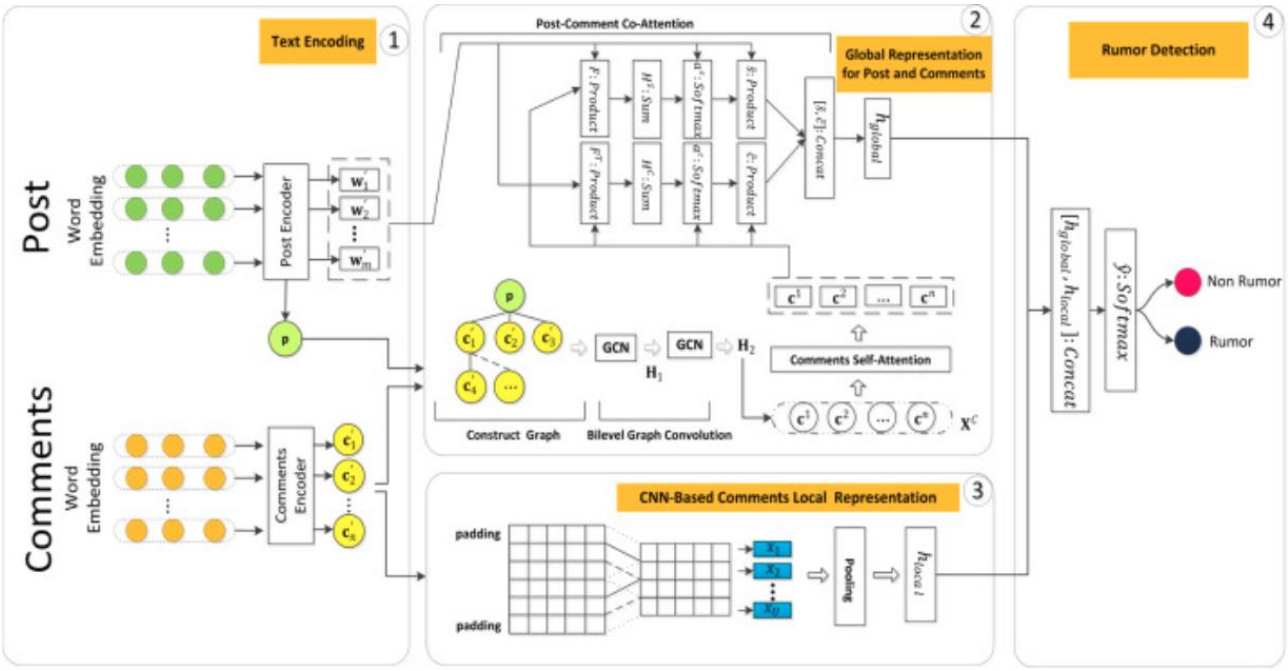
3. Paper proposed model:

The paper PostCom2Dr uses the reply structure of comments on post to detect rumour. When there are lots of user comments then they follow a particular pattern if it is rumour. So, the structure can be used to identify whether the post is rumour or not. The post and comment structures for rumour and non-rumour post is given below:

As we can see, the non-rumour and rumour post follows different kind of comment structures. This structure can be captured using graph. There is also a *topic drift* within comments as time goes on.



The proposed model of the paper first construct graph based on the reply relationship between post and comments. Then a bilevel GCN and self-attention is used to learn the representation of comments based on reply structure. Then post-comment co-attention mechanism is used to learn the mutual selection between post and its comments. The global representation pays attention to the important information of both sides. Topic drift is captured using CNN. The classification uses both global and local representation.



Above model generates Global representation using post and comments structure. Local representation is generated using CNN. Result of both is used to detect rumour.

3.1 Text encoding:

Text encoding is done using two type of encoders: Post encoders and Comments encoder. The post encoder takes the post as input and generates is encoding using LSTM. The Comments encoder also does same but for each comment one encoding is done using LSTM. The result of post encoder and comments encoder is passed to the local and global representation.

3.2 Global representation for post and comments:

The global representation for post and comments will take the encoded post and comments and create the graph from the reply graph using adjacency matrix. This will help model learn the structure of comments on the post. The adjacency matrix is denoted as:

$$\mathbf{a}_{ij} = \begin{cases} 1 & \text{if} \quad \mathbf{e}_{ij} \in \mathbf{E} \\ 0 & \text{otherwise} \end{cases}$$

Then Bi-level GCN is applied on that graph to generate the representation. Here, H1 and H2 are the hidden features of two layer GCN.

$$\mathbf{H}_1 = \tanh\left(\mathbf{A}\mathbf{X}\mathbf{W}_0\right)$$

$$\mathbf{H}_2 = \tanh\left(\mathbf{A}\mathbf{H}_1\mathbf{W}_1\right)$$

The result of Bi-level GCN is input to self-attention as:

$$\mathbf{X}_{att}^{c} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right)\mathbf{V}$$

Note that Q = K = V = X^c and X_{att}^c contains importance of each comment to all comments. This part determines the semantic similarities within the comments. The result of self-attention is given to the co-attention part. Here the intermediate output generated by post encoder is used to take importance of each word from post along with the result of

comments self attention. This will allow model to learn which comments are more relevant to post and which word from post is important for the comments.
First affinity matrix is computed as: It can be seen as the public space of post and comments and through affinity matrix we can learn the attention maps between the words of post and comments:

$$\mathbf{F} = \tanh\left(\mathbf{C}^{\mathbf{T}}\mathbf{W}_{cw}\mathbf{W}'\right)$$

The learnt attention maps between words of post and comments can be shown as:

$$\mathbf{H}^{c} = \tanh\left(\mathbf{W}_{w}\mathbf{W}' + \left(\mathbf{W}_{c}\mathbf{C}\mathbf{F}\right)\right)$$

$$\mathbf{H}^{w} = \tanh\left(\mathbf{W}_{c}\mathbf{C} + \left(\mathbf{W}_{w}\mathbf{W}'\mathbf{F}^{\mathbf{T}}\right)\right)$$

Then softmax function is applied on above results with learnable weights:

$$\mathbf{a}^{w} = \text{softmax}\left(\mathbf{W}_{hw}^{\mathbf{T}}\mathbf{H}^{w}\right)$$

$$\mathbf{a}^{c} = \text{softmax}\left(\mathbf{W}_{hc}^{\mathbf{T}}\mathbf{H}^{c}\right)$$

Here a^w and a^c are the attention probabilities of each word and each comment. The comment and post attention vectors are:

$$\tilde{\mathbf{w}} = \sum_{i=1}^{m} \mathbf{a}_{i}^{w}\mathbf{w}_{i}'$$

$$\tilde{\mathbf{c}} = \sum_{j=1}^{n} \mathbf{a}_{j}^{c}\mathbf{c}^{j}$$

The result of co-attention from above formula shows the global representation. The global feature is represented as $h_{global} = [\tilde{w}, \tilde{c}] \in \mathbb{R}^{1\times(l+m')}$. This is passed for rumour detection.

3.3 Local representation of comments:
The comments may have topic drift. In order to deal with that, the encoded result of comments is used to generate CNN and comments structure is learnt here. The time of comment also helps with detecting the topic drift. CNN helps with this. At a time T number of comments are taken and ReLU function is applied. Then h_{local} is calculated using max pooling operation. Then after pooling layer local representation is generated.

3.4 Rumour detection:
Rumour is detected using local and global representation by concatenating local and global representation. Softmax function is applied to find the final result.

Phase 2:

1. Limitation of proposed model in research paper:

The model shown in paper uses mutual attention to get more post related content from comments and more comments related content from post. This will allow it to eliminate noise and irrelevant information. However, the model failed to select explanatory comments for the results of rumour detection as mentioned by paper. Other than that, the proposed model uses post and comments to detect rumour. Social media contains more information than just post and comments. User information, Hashtags, image, video, etc can also be used to detect the rumour more accurately.

2. Objective of CS 529 phase 2 project:

The Objective of CS 529 phase 2 project is to detect the rumour from news where tweets and comments are associated with news.

3. Explanation of intuitions behind objective:

The model represented in paper uses the post and comments to detect rumour. If the source of post is some news article, then that article is shared on twitter as post. If the source of news is rumour, then posts related to it and comments related to it are also part of that rumour. The pattern followed by the posts and rumours is also same. We can create the global representation and bi-level GCN for complete posts and their comments' structure. Then the co-attention can be applied to post-comments structure and news article. This will allow us to detect the source of rumoured news. This is the motivation for us to detect the rumour from news article.

4. Supporting experimental setup:

Experiment of intuitions is done by making one news article as main post and the twitter post and comments structure is maintained to create bi-level GCN. The news post is sent as main post in the model and twitter posts are considered comments. The word embedding is done using post encoder and the encoded result is the input to the post-comment co-attention. The final state of encoder is given to create GCN with comments. The comments encoder will encode the comments and the final state of each comment is given along with final state of comments. The GCN is applied to this state results. Here also bi-level GCN is performed on the graph. Then the result is given to the comment self-attention to know the relationship between comments. The result of self-attention is then given to post-comment co-attention. This mutual attention finds the relevance between news post, tweets and comments. This was global representation. The local representation applies the CNN on the comments. The global and local representations are used to detect rumour.

5. Produced result over data provided by TA for respective project:

The result of experiment for two data sets Twitter 15-16 and FakeNewsNet is shown below.

Dataset name	Accuracy	F1 score(Class 1)	F1 score(Class 0)
Twitter 15-16	88.429%	0.8511	0.8923
Fake News Net	81.482%	0.8787	0.6081

6. Future work:

Social media contains other information such as image, video, hashtag, expression, etc. These can also be used along with content. So, for the future work, the data related to this can be added to get more accurate model.