Rumour Detection and Veracity Verification

Sukanya Harshvardhan
PES University
PES1201700214
Bangalore, India
sukanya31399@gmail.com

Sirisha Lanka
PES University
PES1201700294
Bangalore, India
siri181.lanka@gmail.com

Prajna Girish
PES University
PES1201701261
Bangalore, India
prajna2310@gmail.com

Prof. Bhaskarjyoti Das
PES University
Bangalore, India
bhaskarjyoti01@gmail.com

t

***Abstract*—Sharing content on social media platforms such as Twitter is one of the most effective ways to target large audiences and spread awareness. However this privilege can be misused when the shared content includes unverified and mistrusted information in the form of rumours. Hence, detecting rumours and assessing their veracity is a key problem to be addressed on such platforms. In the paper, we have proposed a learning mechanism which incorporates analysis of socio-linguistic data and social graph perspectives.**

I. INTRODUCTION

Society has moved online. Although this has been majorly a boon owing to its accessibility and infinite resource power, there is a major downside which needs to be addressed. People nowadays tend to acquire more information from online social media platforms than traditional media channels, and usually believe most of what they see. With the advent of social media it is possible that information posted by a single user receives the attention of millions of users within a few seconds. Since there are no official restrictions on the content of what is being posted on these social media platforms, there tends to be a lot of unverified information online, also known as rumours. [2] The low cost of information exchange on these online platforms allows information to spread quickly and more widely than ever before, and the more a rumour spreads the more people tend to believe it. Hence, their early detection is of utmost importance.

Our focus for this research problem mainly is on Twitter data, considering that it is the primary social media platform used to post breaking news and other current affairs. Further, usage of a social graph would be perfect to understand the data flow on twitter considering that there are fixed data propagation algorithms - retweets, shares, replies etc. The user's behaviour and credibility can be judged based on his/her connections in the social graph, which would further help us determine if the user's post contains unverified information or not. Finally, the language spoken on social media platforms such as Twitter does not abide to the norms of the conventional English language- there is a limit on the number of characters, people use abbreviations and slang, and there might also be spelling errors. This type of language comes under the umbrella of socio-linguistic data, which is what we would be using in this project. Social information flow and socialization behaviours are not only related to information contents and user interests but also to social relationships. Hence, there is a need to integrate all these factors to construct a better learning model for rumours.

The problem at hand can be split into various auxiliary tasks. [5] First, is the rumour detection aspect, where we determine if a tweet's content contains to be verified rather than it just containing an expression. Second, is the stance taken by a user and to determine his/her feelings towards this particular tweet. Finally, is the veracity verification where we determine if the rumour at hand is true, false or unverified.

II. APPLICATIONS

Rumours in itself mean that it is unverified and hence may or may not come from a credible source. During some unexpected sudden events such as a terrorist attack or a natural calamity, people usually tend to refer to social media platforms before conventional media such as news channels because the latter take much more time to publish news owing to the fact that data needs to be first collected and then verified. This delay is usually not desirable especially when people are in a state of panic, and hence their first instinct is to collect whatever information that they can gather from sites such as Twitter. This entire process has been taken advantage of, because the users' primary focus shifts from retaining the veracity of the news to adding extra information to make it more appealing to the public.

The spread of rumours have brought down governments all across the world, destroyed the stock market, and ruined the reputation associated with highly influential individuals. From the examination of Google's search results, it was found that Tweeter's (users on Twitter) and blogs' real-time information mainly consist of "fabricated content, lies, misinterpreta-tion, and unverified events".

For example, during the Covid-19 pandemic that shook the world in 2020, residents of a village in Punjab, India refused to get themselves tested for the same. The reason for the same was the fact that there were rumours being circulated on Twitter and WhatsApp about "Covid being a hoax", and it was all part of an organ harvesting scandal. The public believed that people's organs were being harvested under the guise of diagnosing and treating coronavirus.This news spread like rapid fire and reached almost all the residents of the village, which led to violent protests the public attacking the health workers, including the doctors and nurses. This led to the death rates in the town to skyrocket, but yet the public was headstrong, and was convinced that Covid-19 was not real. Hence, detecting rumours on Twitter and being able to classify whether the information in it is true or not, is an issue that needs to be addressed.

III. PROBLEM STATEMENT

Detecting rumours on online platforms such as Twitter and determining their veracity is a crucial task in preventing the unverified information from propagating across the network.

Over the past few years, various methodologies have been implemented to solve this problem. Multiple ML and Deep Learning approaches have been researched, which analyse the context of a post to classify it as a rumour or non-rumour have been the most common approaches.

Through our research, we propose to build a joint learning model that integrates the socio linguistic data features present in the post along with a social graph perspective. This joint learning model will help include various features such as the content of the post, the propagation pattern as well as features extracted from the social network analysis of the network the post is a present in. A joint learning model with the inclusion of many more features as compared to a single model will help us detect rumours and determine their veracity with more accurate results.

IV. RELATED WORK

A. Content Based Learning

Textual content of a post has been one of the primary attributes which has been utilized for rumour detection. It includes the content in the source post and all user replies following it. [1]In this paper, the model proposed was based on recurrent neural networks (RNN) for learning the hidden representations that capture the variation of contextual information of relevant posts over time.

B. Sequence Based Learning

[4] Veracity verification has been carried out by detecting the stance of each individual tweet, by considering the textual content of the tweet, its timestamp, as well as the sequential conversation structure leading up to the target tweet

C. Propagation Based Learning

Studies have shown that the propagation structures of rumour posts vs non-rumour posts vary and these structures can be analysed to help detect rumours on any social media site. [3] proposes a kernel-based method called Propagation Tree Kernel, which captures high-order patterns differentiating different types of rumors by evaluating the similarities between their propagation tree structures.

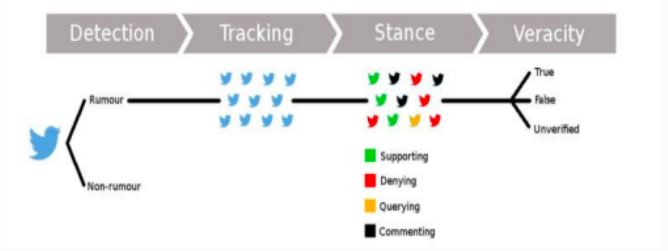
D. Social Graph Based Learning

Rumor detection in streaming social media is a significant but challenging problem. A novel graph-based pattern matching algorithm is described in [6], to detect rumor patterns from streaming social media data. There is special focus on micro cascade motifs that present representative characteristics in an event diffusion network.

E. *Understanding User Stance*

Extensive research has been done using [7] convolutional neural network (CNN) CNN and BERT neural network language models to learn attitude representation for user comments without human annotation via transfer learning based on external data sources for stance classification.

F. *Auxiliary and Main Task*



We have understood that the task [7] involves first detecting if a tweet is a rumour or not. If it has been classified as a rumour, we then collect related relevant tweets- considering that tweets are contextual, and one tweet in isolation will not be able to give us adequate information about the situation. Following this, we understand the user's reactions and feelings towards this tweet- whether they are supporting, denying, querying, or commenting. On the basis of this, we can draw conclusions based on their attitude towards the tweet and their position and connections in a social graph. Finally, we execute the main task of determining the veracity of the tweet.

V. DATASET

Some of the initial studies done on rumor detection can be traced back to 1998, as a patent filed by Microsoft Corporation. This study focused on identifying influential rumormongers through resource usage data using a greedy graph covering algorithm. Since then, detection and evaluation of rumors has come a long way with the help of Natural Language Processing and deep-learning.

We have used different datasets for different subtasks. Although the structures of all are the same, they vary slightly in terms of the features and qualities that they capture and represent. The data is organised as undermentioned: The dataset consists of 9 events, each of which is divided into two main subfolders: the rumours and the non-rumours. These folders in turn consist of tweet-id folders. These tweet ids correlate to the tweet id of the associated source tweet. Each tweet id folder has four main components: the source tweet folder, the replies folder, the structure or flow of conversation in json (structure.json) and the labels associated with that particular conversation (such as veracity) in json (annotations.json).

As long as a dataset with a similar structure and organisation is used, the model should perform in the same manner. We have decided to use one of the most extensive datasets, the Pheme dataset for the purpose of rumour detection ie.determining whether a tweet contains information that is verified or not. Many variations of this dataset exist, but we use the Pheme dataset for Rumor detection and Veracity Verification. It is a popular research intended compilation of Twitter rumours and non-rumours posted during breaking news.
However, for the purpose of the rest of the subtasks, namely stance classification and rumour veracity verification, we use the SemEval2019 Dataset. This consists of Twitter as well as Reddit data. It shares the same structure as Pheme, but captures more features about the user - such as the number of followers and the number of people he/she is following, user verification, number of retweets etc. This helps determine the user credibility, which is an important determining factor for whether a tweet is a rumour or not.

VI. MODULES USED

- os module (*inbuilt*): provides a portable way of using system functionality. It is mostly used to parse and process the dataset.
- json module (*ver 2.0.9*): used to process data in json format.
- scikit-learn module (*ver 0.23.2*): machine learning library that supports supervised and unsupervised learning.
- nltk (*ver 3.5*): platform to build python programs to work with human language data.
- transformers (*ver 3.5.1*): library with NLP-oriented architectures of pre-trained models
- pytorch (*ver 1.7.0*): open-source library that is based on the Torch machine learning library.
- numpy (*ver 1.19.0*): module that will assist in working with large numerical data.
- re (*ver 2.2.1*): helps evaluate regular expressions in textual data.
- Word2vec module to capture the content of a tweet
- torchmoji to capture the emotions associated with a particular tweet

VII. SYSTEM DESIGN

Social Media sites such as Twitter and Reddit are usually the first ones to get first hand accounts of breaking news, because conventional media sources take time to publish information because it needs to be verified before broadcasting it. Hence, most of the population relies heavily on social media sites, mainy Twitter, for an instant recount of drastic events that have taken place, seconds after they have occured.

Initially, right after an event occurs, there are numerous tweets about this particular event, which can all be tagged because they will all use a common hashtag, something that describes the event and its happenings.

For the purpose of rumour detection, we have used a BERT transformer. Our model is given all the related tweets and then performs various forms of data cleaning and data preprocessing. First off, all the Emoji characters are removed, making the text more easy to comprehend and understand. Next, the words in the Tweet are embedded using BERT which performs deep bidirectional embedding, trying to capture all token and position related information to produce a real-valued vector. The elements of these vectors serve as parameters to our model.

Finally, the embedded vectors in the Tweet are stored in a database, for easy data retrieval and access. The tweets are then fed into our transformer model, as part of the training data and our model learns the optimal weights for different parameters through backpropagation. These weights are fine-tuned further using Adam optimization.

The next stage of our project is where the user and his/her input comes into play. A user wants to verify if a given tweet with a particular tweet ID contains information that is true or not. The user inputs the tweet into our model. The same procedure of Emoji character removal and word vectorisation is followed, and done as part of the data preprocessing phase. Following this, our model predicts whether the content of the tweet is agreeing/ denying/questioning a topic, rather than stating a topic.

Next, we have the graph based model for rumour detection and veracity verification. For every tweet- we extract 5 features from it, namely:

- Number of followers of the user
- Number of accounts the user follows
- Whether the user is verified on Twitter or not
- Tweet's favourite count
- Tweet's retweet count

An important feature that needs to be captured is the propagation structure of the tweet- how exactly it travels through the social network of Twitter and how it moves from one user to another. This can be captured by the Tweet's *favourite* and *retweet* count. For every tweet, we create a dictionary for all 5 of these features, with the key as the feature name and the corresponding value of that feature. This dictionary and all associated features and then vectoried using python modules. Following this, the obtained vector is being fed into a decision tree, which does the task of classifying it as a rumour or not. The obtained model gives us an accuracy of 73%.

Finally, we combine both the rumour detection models by assigning a weightage to each model on the basis of its accuracy. We also look into the confidence scores offered by both of the models and then make an informed decision on what has a higher weightage. After combining both models, we have successfully obtained an accuracy of 93%.

On the basis of this, it decides whether the tweet is a rumour or not. If not, the tweet is dropped. Based on content on individual tweets and their graph perspectives, we classify them as rumors or not. With this, we are able to detect rumors.

The next phase of our project deals with the stance classification, which classifies whether a given tweet is supporting, denying, querying or commenting the given source tweet. The stance of a tweet is being classified with the help of a Bi-LSTM and uses factors such as cosine similarity and the number of negative words/ punctuations to determine how similar or different the reply tweet is from the source.

In addition to this, we are using the torchmoji vector in python for better results. With the help of this module, emoticons are assigned to a particular tweet, based on the emotion present in the content of hte tweet. The emotion associated with the majority of the emoticons in the tweet is the emotion assigned to the tweet in itself. A combination of this module and the Bi-LSTM gives us an accuracy of 78%.

Finally, is the veracity verification aspect where we finally produce the outcome of whether a Tweet is true, false or unverified. On the basis of the stance taken by all of the source tweet's replies, we use a Bi-LSTM to finally predict the outcome. This final model provides us with an accuracy of 92%. The model has three outcomes - True, False or unverified.

VIII. CONCLUSION AND FUTURE WORK

Through this project, we aimed to analyse rumours on the basis of two of their important features: Linguistic Content and Social Graphs.

For the social graph perspective, we selected extracted features from the dataset that help gauge the presence of an individual in a social network.

Then, we developed a model that could learn rumours over social graph related information like user relationships, rumour propagation patterns and more. Further, we combined the loss functions of the linguistic based model and graph based model to generate a final model. This newly constructed model can classify a rumour using its two facets: sentiment relayed through text and behaviour identified in social graphs. Once combined, we performed a final analysis on this model with the firm belief that it shall surpass other existent models.

As observable from the decreasing training and validation loss, as well as the improving confusion matrix values through all the epochs, we can confidently claim that this model works efficiently.

It is very important to detect and verify rumours as soon as possible so as to minimize the potential damage that it can cause. Our current model does precisely that and we hope to further fine-tune it in the future to achieve higher efficiencies and also extend it to different social media platforms, such as Reddit, Facebook and Instagram as well.

VIII. REFERENCES

[1]MA, Jing; GAO, Wei; MITRA, Prasenjit; KWON, Sejeong; JANSEN, Bernard J.; WONG, Kam-Fai; and CHA, Meeyoung. Detecting rumors from microblogs with recurrent neural networks. (2016). *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*. 3818-3824. Research Collection School Of Information Systems.

[2] Ahsan, Mohammad & Kumari, Madhu & Sharma, T. (2019). Rumors detection, verification and controlling mechanisms in online social networks: A survey.

[3] MA, Jing; GAO, Wei; and WONG, Kam-Fai. Detect rumors in microblog posts using propagation structure via kernel learning. (2017). Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver, Canada, 2017 July 30 - August 4. 708-717. Research Collection School Of Information Systems.

[4] L. Poddar, W. Hsu, M. L. Lee and S. Subramaniyam, "Predicting Stances in Twitter Conversations for Detecting Veracity of Rumors: A Neural Approach," 2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI), Volos, 2018, pp. 65-72, doi: 10.1109/ICTAI.2018.00021.

[5] Pavithra C P , Shibily Joseph, Detection and Verification of Rumour in Social Media: A Survey

[6] Shihan Wang, Takao Terano, Detecting rumor patterns in streaming social media

[7]Tian L., Zhang X., Wang Y., Liu H. (2020) Early Detection of Rumours on Twitter via Stance Transfer Learning. In: Jose J. et al. (eds) Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, vol 12035. Springer, Cham. https://doi.org/10.1007/978-3-030-45439-5_38

[8]Zubiaga, A.; Liakata, M.; Procter, R.: Exploiting context for rumour detection in social media. In: Ciampaglia, G.L., Mashhadi, A., Yasseri, T. (eds.) Social Informatics, pp. 109–123. Springer, Cham (2017)

[9] Alsaeedi, A., Al-Sarem, M. Detecting Rumors on Social Media Based on a CNN Deep Learning Technique. *Arab J Sci Eng* (2020). https://doi.org/10.1007/s13369-020-04839-2