Bi-Directional Recurrent Neural Ordinary Differential Equations for Social Media Text Classification

Maunika Tamire, Srinivas Anumasa, P.K. Srijith
Computer Science and Engineering
Indian Institute of Technology Hyderabad, India
cs18mds11026@iith.ac.in, cs16resch11004@iith.ac.in, srijith@cse.iith.ac.in

Abstract

Classification of posts in social media such as Twitter is difficult due to the noisy and short nature of texts. Sequence classification models based on recurrent neural networks (RNN) are popular for classifying posts that are sequential in nature. RNNs assume the hidden representation dynamics to evolve in a discrete manner and do not consider the exact time of the posting. In this work, we propose to use recurrent neural ordinary differential equations (RNODE) for social media post classification which consider the time of posting and allow the computation of hidden representation to evolve in a time-sensitive continuous manner. In addition, we propose a novel model, Bi-directional RNODE (Bi-RNODE), which can consider the information flow in both the forward and backward directions of posting times to predict the post label. Our experiments demonstrate that RNODE and Bi-RNODE are effective for the problem of stance classification of rumours in social media.

1 Introduction

Information disseminated in social media such as Twitter can be useful for addressing several real-world problems like rumour detection, disaster management, and opinion mining. Most of these problems involve classifying social media posts into different categories based on their textual content. For example, classifying the veracity of tweets as False, True, or unverified allows one to debunk the rumours evolving in social media (Zubiaga et al., 2018a). However, social media text is extremely noisy with informal grammar, typographical errors, and irregular vocabulary. In addition, the character limit (240 characters) imposed by social media such as Twitter make it even harder to perform text classification.

Social media text classification, such as rumour stance classification[1] (Qazvinian et al., 2011; Zubiaga et al., 2016; Lukasik et al., 2019) can be addressed effectively using sequence labelling models such as long short term memory (LSTM) networks (Zubiaga et al., 2016; Augenstein et al., 2016; Kochkina et al., 2017; Zubiaga et al., 2018b,a; Dey et al., 2018; Liu et al., 2019; Tian et al., 2020). Though they consider the sequential nature of tweets, they ignore the temporal aspects associated with the tweets. The time gap between tweets varies a lot and LSTMs ignore this irregularity in tweet occurrences. They are discrete state space models where hidden representation changes from one tweet to another without considering the time difference between the tweets. Considering the exact times at which tweets occur can play an important role in determining the label. If the time gap between tweets is large, then the corresponding labels may not influence each other but can have a very high influence if they are closer.

We propose to use recurrent neural ordinary differential equations (RNODE) (Rubanova et al., 2019) and developed a novel approach bidirectional RNODE (Bi-RNODE), which can naturally consider the temporal information to perform time sensitive classification of social media posts. NODE (Chen et al., 2018) is a continuous depth deep learning model that performs transformation of feature vectors in a continuous manner using ordinary differential equation solvers. NODEs bring parameter efficiency and address model selection in deep learning to a great extent. RNODE generalizes RNN by extending NODE for time-series data by considering temporal information associated with the sequential data. Hidden representations are changed continuously by considering the temporal information.

We propose to use RNODE for the task of sequence labeling of posts, which considers arrival times of the posts for updating hidden representa-

[1] Rumour stance classification helps to identify the veracity of a rumour post by classifying the reply tweets into different stance classes such as Support, Deny, Question, Comment

Proceedings of the 2nd Workshop on Deriving Insights from User-Generated Text, pages 20 - 24
May 27, 2022 ©2022 Association for Computational Linguistics

tions and for classifying the post. In addition, we propose a novel model, Bi-RNODE, which considers not only information from the past but also from the future in predicting the label of the post. Here, continuously evolving hidden representations in the forward and backward directions in time are combined and used to predict the post label. We show the effectiveness of the proposed models on the rumour stance classification problem in Twitter using the RumourEval-2019 (Derczynski et al., 2019) dataset. We found RNODE and Bi-RNODE can improve the social media text classification by effectively making use of the temporal information and is better than LSTMs and gated recurrent units (GRU) with temporal features.

2 Background

We consider the problem of classifying social media posts into different classes. Let $\mathcal{D}$ be a collection of N posts, $\mathcal{D} = \{p_i\}_{i=1}^{N}$. Each post p_i is assumed to be a tuple containing information such as textual and contextual features $\mathbf{x}_i$, time of the post t_i and the label associated with the post y_i, thus $p_i = \{(\mathbf{x}_i, t_i, y_i)\}$. Our aim is to develop a sequence classification model which considers the temporal information t_i along with $\mathbf{x}_i$ for classifying a social media post. In particular, we consider the rumour stance classification problem in Twitter where one classifies tweets into Support, Query, Deny, and Comment class, thus $y_i \in$ Y={Support, Query, Deny, Comment}.

2.1 Neural Ordinary Differential Equations

NODE were introduced as a continuous depth alternative to Residual Networks (ResNets) (He et al., 2016). ResNets uses skip connections to avoid vanishing gradient problems when networks grow deeper. Residual block output is computed as $\mathbf{h}_{t+1} = \mathbf{h}_t + f(\mathbf{h}_t, \theta_t)$, where $f()$ is a neural network (NN) parameterized by θ_t and $\mathbf{h}_t$ representing the hidden representation at depth t. This update is similar to a step in Euler numerical technique used for solving ordinary differential equations (ODE) $\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t), t, \theta)$. The sequence of residual block operations in ResNets can be seen as a solution to this ODE. Consequently, NODEs can be interpreted as a continuous equivalent of ResNets modeling the evolution of hidden representations$\mathbf{h}(t)$ over time.

For solving ODE, one can use fixed step-size numerical techniques such as Euler, Runge-Kutta or adaptive step-size methods like Dopri5(Dormand and Prince, 1980). Solving an ODE requires one to specify an initial value $\mathbf{h}(0)$ (input $\mathbf{x}$ or its transformation) and can compute the value at t using an ODE solver $ODESolverCompute(f_\theta, \mathbf{h}(0), 0, t)$. An ODE is solved until some end-time T to obtain the final hidden representation $\mathbf{h}(T)$ which is used to predict class labels $\hat{y}$. For classification problems, cross-entropy loss is used and parameters are learnt through adjoint sensitivity method (Zhuang et al., 2020; Chen et al., 2018) which provides efficient back-propagation and gradient computations.

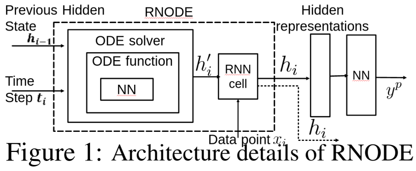


Figure 1: Architecture details of RNODE

3 Bi-Directional Recurrent NODE

LSTMs are popular for sequence classification but only considers the sequential nature of the data and ignore the temporal features associated with the data in its standard setting. As the posts occur in irregular intervals of time, the nature of a new post will be influenced by the recent posts, influence will be inversely proportional to the time gap. In these situations, it will be beneficial to use a model where the number of transformations depend on the time gap.

We propose to use RNODE which considers the arrival time and accordingly the hidden representations are transformed across time. In RNODE, the transformation of a hidden representation $\mathbf{h}(t_{i-1})$ at time t_{i-1} to $\mathbf{h}(t_i)$ at time t_i is governed by an ODE parameterized by a NN $f()$. Unlike standard LSTMs where $\mathbf{h}(t_i)$ is obtained from $\mathbf{h}(t_{i-1})$ as a single NN transformation, RNODE first obtains a hidden representation $\mathbf{h}'(t_i)$ as a solution to an ODE at time t_i with initial value $\mathbf{h}(t_{i-1})$. The number of update steps in the numerical technique used to solve this ODE depends on the time gap $t_i - t_{i-1}$ between the consecutive posts. The hidden representation $\mathbf{h}'(t_i)$ and input post $\mathbf{x}_i$ at time t_i are passed through neural network transformation (RNNCell()) to obtain final hidden representation $\mathbf{h}(t_i)$, i.e., $\mathbf{h}(t_i)$ = RNNCell($\mathbf{h}'(t_i), \mathbf{x}_i$). The process is repeated for every element $(\mathbf{x}_i, t_i)$ in the sequence. The hidden representations associated with the elements in the sequence are then passed to a neural network (NN()) to obtain the post labels. Using standard cross-entropy loss, the parameters of the models are learnt through backpropagation. Figure 1 provides the detailed architecture of the

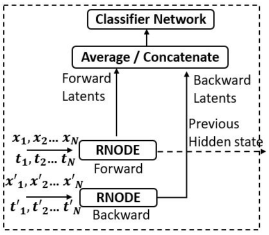


Figure 2: Bi-RNODE Architecture

RNODE model.

Bi-directional RNNs (Schuster and Paliwal, 1997) such as Bi-LSTMS (Graves et al., 2013) were proven to be successful in many sequence labeling tasks in natural language processing such as POS tagging (Huang et al., 2015). They use the information from the past and future to predict the label while standard LSTMs consider only information from the past. We propose a Bi-RNODE model, which uses the sequence of input observations from past and from the future to predict the post label at any time t. It assumes the hidden representation dynamics are influenced not only by the past posts but also by the futures posts. Unlike Bi-LSTMs, Bi-RNODE considers the exact time of the posts and their inter-arrival times in determining the transformations in the hidden representations. Bi-RNODE consists of two RNODE blocks, one performing transformations in the forward direction (in the order of posting times) and the other in the backward direction. The hidden representations H and H_b computed by forward and backward RNODE respectively are aggregated either by concatenation or by averaging appropriately to obtain a final hidden representation and is passed through a NN to obtain the post labels. Bi-RNODE is useful when a sequence of posts with their time of occurrence needs to be classified together.

Figure 2 provides an overview of Bi-RNODE model for post classification. For Bi-RNODE, an extra neural network $f_{\theta'}()$ is required to compute hidden representations $\mathbf{h}_b(t'_i)$ in the backward direction. Training in Bi-RNODE is done in a similar manner to RNODE, with cross-entropy loss and back-propagation to estimate parameters.

4 Experiments

To demonstrate the effectiveness of the proposed approaches, we consider the stance classification problem in Twitter and RumourEval-2019 (Derczynski et al., 2019) data set. This Twitter data set consists of rumours associated with eight events. Each event has a collection of tweets labelled with one of the four labels - Support, Query, Deny and Comment. We picked four major events Charliehebdo, Ferguson, Ottawashooting and Sydneysiege (each with approximately 1000 tweets per event) from RumourEval-2019 to perform experiments.

Features : For dataset preparation, each data point $\mathbf{x}_i$ associated with a Tweet includes text embedding, retweet count, favourites count, punctuation features, negative and positive word count, presence of hashtags, user mentions, URLs etc. obtained from the tweet. The text embedding of the tweet is obtained by concatenating the word embeddings [2] . Each tweet timestamp is converted to epoch time and Min-Max normalization is applied over the time stamps associated with each event to keep the duration of the event in the interval $[0, 1]$.

4.1 Experimental setup

We conducted experiments to predict the stance of social media posts propagating in *seen events* and *unseen events*.

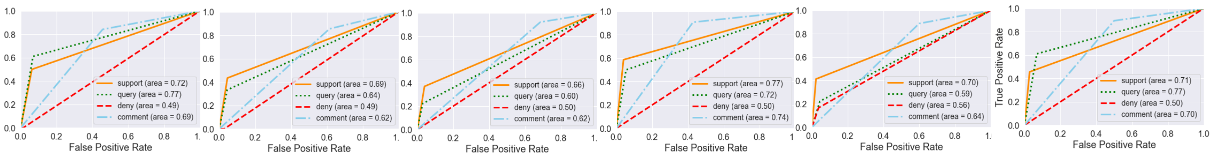
-Seen Event Here we train, validate and test on tweets of the same event. Each event data is split 60:20:20 ratio in sequence of time. This setup helps in predicting the stance of unseen tweets of the same event.

-Unseen Event: This setup helps in evaluating performance on an *unseen event* and training on a larger dataset. Here, training and validation data are formed using data from 3 events and testing is done on the 4^{th} event. Last 20% of the training data (after ordering based on time) are set aside for validation. During training, mini-batches are formed only from the tweets belonging to the same event.

Baselines: We compared results of our proposed RNODE and Bi-RNODE models with RNN based baselines such LSTM (Kochkina et al., 2017), Bi-LSTM (Augenstein et al., 2016), GRU (Cho et al., 2014), Bi-GRU, and Majority (labelling with most frequent class) baseline models. We also use a variant of LSTM baseline considering temporal information (Zubiaga et al., 2018b), LSTM-timeGap where the time gap of consecutive data points is included as part of the input data.

Evaluation Metrics: We consider the standard evaluation metrics such as precision, recall, F1 and in addition the AUC score to account for the data imbalance. We consider a weighted average of the

[2] Using pre-trained word2vec vectors which are trained on Google News dataset: https://code.google.com/p/word2vec, each word is represented as an embedding of size 15.



(a) RNODE (b) LSTM (c) GRU (d) Bi-RNODE (e) Bi-LSTM (f) Bi-GRU

Figure 3: ROC curves of different models trained on sydneysiege event for *seen event* experimental setup. Bi-RNODE exhibits better AUC and class separability overall classes.

Model	Charliehebdo				Ferguson				Ottawashooting			
	AUC	**F1**	**Recall**	**Precision**	**AUC**	**F1**	**Recall**	**Precision**	**AUC**	**F1**	**Recall**	**Precision**
RNODE	0.665	0.653	0.674	**0.658**	**0.600**	0.591	0.659	0.598	0.638	0.654	**0.692**	**0.670**
	0.638	0.672	0.700	**0.721**	**0.618**	0.632	0.677	**0.640**	**0.659**	**0.651**	**0.703**	0.642
Bi-RNODE	**0.696**	**0.659**	**0.693**	0.629	0.595	0.599	**0.673**	**0.641**	**0.669**	**0.667**	**0.692**	0.658
	0.651	**0.697**	**0.737**	0.690	0.615	**0.643**	**0.695**	0.635	0.652	0.624	0.662	0.618
Bi-LSTM	0.628	0.625	0.679	0.609	0.563	0.599	0.650	0.614	0.622	0.627	0.654	0.622
	0.662	0.690	0.717	0.671	0.603	0.623	0.667	0.600	0.650	0.637	0.686	0.622
Bi-GRU	0.654	0.643	0.660	0.641	0.588	0.571	0.631	0.625	0.640	0.651	0.686	0.644
	0.656	0.690	0.724	0.682	0.613	0.634	0.678	0.611	0.648	0.636	0.683	0.610
LSTM	0.625	0.600	0.637	0.637	0.567	**0.602**	0.650	0.611	0.605	0.609	0.635	0.603
	0.645	0.690	0.728	0.686	0.602	0.611	0.631	0.603	0.630	0.626	0.680	0.627
GRU	0.616	0.610	0.647	0.623	0.578	0.588	0.664	0.631	0.591	0.539	0.513	0.574
	0.682	0.695	0.713	0.686	0.614	0.640	0.687	0.623	0.638	0.632	0.683	0.618
LSTM-timeGap	0.638	0.631	0.679	0.605	0.565	0.581	0.627	0.590	0.625	0.640	0.679	0.650
	0.652	0.695	0.732	0.696	0.604	0.625	0.673	0.633	0.638	0.638	0.683	**0.651**
Majority	0.500	0.456	0.605	0.366	0.500	0.518	0.654	0.428	0.500	0.485	0.628	0.395
	0.500	0.542	0.673	0.453	0.500	0.528	0.662	0.439	0.500	0.467	0.614	0.377

Table 1: Performance of all the models on RumourEval-2019 (Derczynski et al., 2019) dataset. First and second rows of each model represents *seen event* and *unseen event* experiment results respectively.

evaluation metrics to compare the performance of models.

Hyperparameters: All the models are trained for 50 epochs with 0.01 learning rate, Adam optimizer, dropout(0.2) regularizer, batchsize of 50, hidden representation size of 64 and cross entropy as the loss function. Different hyperparameters like neural network layers (1, 2), numerical methods (Euler, RK4, Dopri5 for RNODE and Bi-RNODE) and aggregation strategy (concatenation or averaging for Bi-LSTM Bi-GRU and Bi-RNODE) are used for all the models and the best configuration is selected from the validation data for different experimental setups and train/test data splits.

4.2 Results and Analysis

The results of *seen event* and *unseen event* experiment setup can be found in Table 1, where the first and second rows for each model provides results on *seen event* and *unseen event* respectively. We can observe from Table 1 that for both *seen event* and *unseen event* experiment setup, RNODE and Bi-RNODE models performed better than the baseline models in general for all the 3 events[3]. In particular for the *seen event* setup, Bi-RNODE gives the best result outperforming RNODE and other models for most of the data sets and measures. Under *seen event* experiment on Syndneysiege event, we plot the ROC curve for all the models in Figure 3. We can observe that AUC for Figures 3(a) and 3(e) corresponding to RNODE and Bi-RNODE respectively are higher than LSTM, GRU, Bi-LSTM , and Bi-GRU.

5 Conclusion

We proposed RNODE, Bi-RNODE models for sequence classification of social media posts. These models consider temporal information of the posts and hidden representation are evolved as solution to ODE. Through experiments, we show these models perform better than LSTMs on rumour stance classification problem in Twitter

[3] Due to space constraint, Table 1 presents results for 3 events, Syndneysiege results in Figure 3.

References

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111.

Leon Derczynski, Genevieve Gorrell, Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Elena Kochkina. 2019. Rumoureval 2019 data.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *Advances in Information Retrieval*, pages 529–536.

John R Dormand and Peter J Prince. 1980. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Zhiheng Huang, W. Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *ArXiv abs/1508.01991*.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 task 8: Sequential approach to rumour stance classification with branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 475–480.

Y. Liu, X. Jin, and H. Shen. 2019. Towards early identification of online rumors based on long short-term memory networks. *Inf. Process. Manag.*, 56:1457–1467.

Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2019. Gaussian processes for rumour stance classification in social media. *ACM Trans. Inf. Syst.*, 37(2).

Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599.

Yulia Rubanova, Ricky TQ Chen, and David Duvenaud. 2019. Latent odes for irregularly-sampled time series. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5320–5330.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Lin Tian, Xiuzhen Zhang, Yan Wang, and Huan Liu. 2020. Early detection of rumours on twitter via stance transfer learning. In *European Conference on Information Retrieval*, pages 575–588. Springer.

Juntang Zhuang, Nicha Dvornek, Xiaoxiao Li, Sekhar Tatikonda, Xenophon Papademetris, and James Duncan. 2020. Adaptive checkpoint adjoint method for gradient estimation in neural ode. In *International Conference on Machine Learning*, pages 11639–11649. PMLR.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018a. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, and Michal Lukasik. 2016. Stance classification in rumours as a sequential task exploiting the tree structure of social media conversations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2438–2448.

Arkaitz Zubiaga, Elena Kochkina, Maria Liakata, Rob Procter, Michal Lukasik, Kalina Bontcheva, Trevor Cohn, and Isabelle Augenstein. 2018b. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290.