Technology Trend Prediction from Social Media using Long Short Term Memory Network

Mrs. Yesha Mehta
Department of Computer Science
Shree Ramkrishna Institute of Computer Education and
Applied Sciences Surat, India

Dr. Kalpesh Lad
Department of Computer Science
Shrimad Rajchandra Institute of Management and Computer
Application Bardoli, India

Dr. Sanjay Buch
Ex. Prof,
Department of Computer Engineering and Information Technology,
CGPIT, Bardoli,India

Abstract— The present study aims to develop a deep learning based data analysis model which would act as a framework for efficient implementation of Social Media Analysis (SMA). This framework combines four process for data analysis i) Data collection, ii) Pre-processing, iii) Technology Classification and iv) Technology Trend prediction. The study uses hybrid model of Deep Feed Forward Neural Network (DNN) and Long Short Term Memory (LSTM) network. The technology and tools learned by people becomes obsolete in short period of time. As I.T. industry required frequent upgrades in knowledge and new technologies are being released, it is essential to track and know new upcoming technology trend of the field. To achieve this aim deep learning model is developed to identify upcoming technologies from social media threads. This paper presents LSTM trend prediction model to predict technology trends from unstructured text content of social media sources. The proposed method ensembles classification and regression process in single architecture. First, it uses deep learning algorithm to build a classifier to correctly predict the technology topic of a discussion thread from its description. After technology identification, how frequently that technology is discussed with respect to time is calculated to generate temporal series of frequencies. The Long Short Term Memory (LSTM) network is combined with Deep Feed Forward Neural Network model for processing temporal topic sequence and frequencies recursively to predict technology trends from of social posts generated on social platforms.

Keywords— Technology Trend Prediction, Social Media, Deep Feed Forward Neural Network, Long Short Term Memory Network

I. INTRODUCTION

The distillation of knowledge from unstructured social big data is, an extremely challenging tasks. Existing Social Media Analytics approaches includes machine learning and deep learning based model which are having limitations in identifying patterns and trends from unstructured content. Deep Learning algorithms extract high-level, complex abstractions as data representations through a hierarchical learning process. Deep Feed Forward Neural Network model learns underlying representations from data itself. Deep Neural Network model outperforms machine learning models in the task of classification and are considered as good classifiers, yet main component of Deep Neural Network is feed forward neural network which is not designed for time dependant data. Deep Neural Network is not suitable for time series based trend prediction problem where network requires context of previous observations.

Deep learning models do not have any understanding of their input, at least not in any human sense. There is fundamental differences between the straightforward geometric morphing from input to output that deep learning models do, and the way that humans think and learn. Humans can easily identify trending topics/things/places by what is being talked most and discussed most. In contrast, machine/model cannot perform this task in a single step. To achieve this goal, computing model should be able to identify technology from social threads, its relevance with time and requires mathematical quantification. In the presented work, process of trend analysis is decomposed into Classification and Tagging, Temporal Mapping and Trend Prediction.

The key characteristic of social platform is the continuous generation of content which leads to derivation of new knowledge out of it. Social platforms are dynamic in nature and they can be considered as new type of information resources for future trend predictions with application of data analytics techniques. The discussion threads and content available on the websites such as Twitter, Facebook and other web portals includes latest information and people's view on different subjects and topics. Using social media data to solve domain-specific problem is challenging due to complexity of the domain, lack of context, colloquial nature of language. Considering the aspect of problem-solving, the content diversity and data volume inherent in social media create significant practical challenges for extracting relevant information as it is similar to searching for a needle in a haystack. Identifying the "Emerging Technology Trends" from social media is an open challenge and appealing work which leads to the exploration of presented research problem. This type of research work can help stakeholders of education community including students, academic institutions, management, university etc. to support their decision making process in areas of employment, career path selection and curriculum designing process. The uniqueness of social media data calls for novel data mining techniques that can effectively handles user generated content. The motivation and idea behind

this research is to design a concept for processing open data available on social platforms to predict trends of emerging technologies of I.T. field.

II. NEED OF LONG SHORT TERM MEMORY NETWORK FOR TECHNOLOGY TREND PREDICTION

Following approaches are taken by researchers to process textual content of social platforms where limitations and data processing challenges are discussed in this section.

A. Natural Language Processing and Semantic Analysis based approach:

As reported in literature, NLP and Statistical Models are adopted previously to analyse and understand text. Statistical NLP relies on language models (Manning & SchÜtze, 1999; Hofmann, 1999; Nigam, McCallum, Thrun, & Mitchell, 2000) based on Machine Learning algorithms such as Expectation Maximization (Nigam et al., 2000), Conditional Random Fields (Lafferty, McCallum, & Pereira, 2001), and Support Vector Machines (Joachims, 2002). By feeding a large training corpus of annotated texts to a machine-learning algorithm, it is possible for the system to learn keywords, punctuation, and word co-occurrence frequencies. Statistical methods are semantically weak which identifies only obvious keywords and model will have less predictive value. As a result, statistical text classifiers only work with acceptable accuracy when given a sufficiently large text input. While these methods are able to classify text on the page-or-paragraph level, they do not work well on smaller text units such as sentences or clauses. The social posts have certain character limits and facts are presented in short text only. As social posts are having less words and limited content in it, to classify them better approach is requireed.

B. Machine Learning based Methods

In Data Analytics, ML based methodologies are used to devise and generate complex algorithms and models which lend themselves to a prediction. Traditional ML model performs feature extraction and model construction in a separated manner. The handcrafted features are firstly extracted by transforming raw data into a form of statistical quantity, frequency, and time-frequency to take the representative information before applying predictive models. The performance of the constructed model not only relies on the optimization of adopted algorithms but also is heavily affected by the handcrafted features. The feature extraction and selection are time-consuming which requires the complex data pre-processing methods depending upon the type data. Machine learning models requires manual feature engineering for individual data sources. In case of heterogeneous social data sources, pre-processing of data increases the model complexity.

C. Deep Learning based Methods

In deep learning based methods, features are learned by transforming data into abstract representations. Recently, deep learning approaches have gained attention from the research community and industry for their ability to automatically learn optimal feature representation for a given task, while claiming state-of-the-art performance in many tasks in computer vision, speech recognition and natural language processing. It has been observed that with large amount of training data, deep neural networks are able to efficiently map the raw input of text to a low-dimensional vector representation, which preserves important syntactic and semantic aspects of the input text. Deep

learning models outperforms other models in task of classification, but such model can handle one problem at a time. Either classification or regression type of problems are handled at a time for data analysis. Deep Neural Network model generates only single output y for series of input. Such network does not support multiple outputs/series based on many given inputs. Main component of DNN is feed forward neural network which is not designed for time dependant data. Problem of trend analysis and prediction requires to process data which is having time parameter or time based sequence pattern. Each input shown to DNN is processed independently, with no state kept in between inputs. With such networks, in order to process sequence or temporal series of data points, entire sequence needs to be presented to network at once which is difficult in case of large amount of data. DNN is not suitable for time series or trend prediction problem where network requires context of previous observations as it suffers from vanishing gradient [5] issue with more hidden layers, where adding more layers to network does not work and model training will become still. DNN is sufficient for single task of classification, with less computational cost and less training time but cannot work with temporal data.

LSTM networks are designed for time series based prediction. While feed-forward networks map many input to only one output, recurrent networks can map one input to many output, many to many (translation), or many to one (classifying a voice). LSTM networks carries information across time steps, and they require more processing resources and time to train the model. Thus, use of LSTM is preferable for sequential and time based data only.

D. Deep Feed Forward and LSTM Model

In the presented work, the deep learning based data analysis model is designed to perform following task of social media analysis and technology trend prediction in single end-to-end architecture.

More formally, given a series of social posts S= {s1 ,s2 , . . . ,sn} which are created on time stamp TS ={ts1 ,ts2 , . . . ,tsT } where T $\in$ R, we aim at predicting a future technology trends by generating time based sequences from unstructured textual content of social posts. To predict T_{t+h} where h is the desirable horizon ahead of the current time stamp, social posts are classified in to technology topics by deep feed forward network and new dataset S = {{y1 ,y2 , ... ,yT },{ts1 ,ts2 , . . . ,tsT }} is generated where y1, y2 are predicted technologies from content and ts1, ts2 are timestamp. To predict the value of the next time stamp T_{t+h+1}, LSTM network is combined with deep feed forward neural network.

This research work focuses on the development of method where deep learning model is implemented for trend prediction from text content of social media sources. A major challenge in developing such deep learning model is to manage the data representation and transform data such that network models can be trained to generate expected output. DL models are capable to predict future outcome from data, but learning happens from the data itself. This model is developed to work with unstructured content without any quantitative values. The data representation from which neural network learn to predict pattern is managed at each step such that models are trained to predict short and long term patterns together in single architecture.

III. MODEL ARCHITECTURE

Following diagram presents the process flow of model generation and implementation where three major process are designed and developed to transform content for integrating Deep Feed Forward and LSTM in a single architecture design.
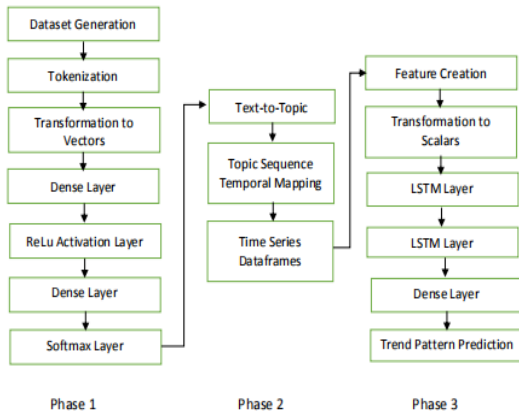

Fig. 1 Hybrid Neural Network Architecture for Technology Trend Prediction

Phase 1: In first phase, model attempts to read the content of social post and identify the technology topic to tag it. This aim is achieved by training deep neural network model to classify social post to its relevant topic. Following figure 2 presents the process of technology classification by applying deep feed-forward neural network for classifying any text into class of relevant technology.
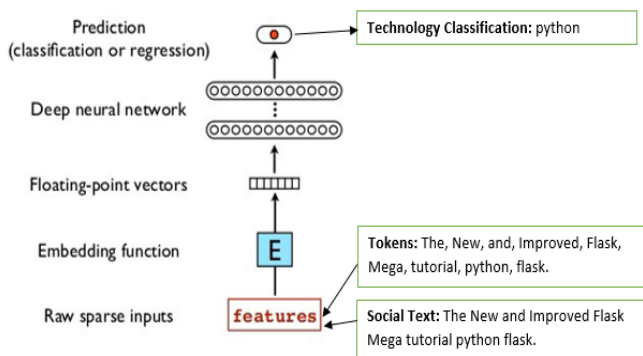

Fig 2. Deep Feed Forward Network for Technology Topic Identification

Phase 2: After technology topic identification, the topic sequence and time mapping needs to be performed for identifying the past trend patterns of posted topics. The datasets which contains details of date-time and topic, will serve as input to another network for future trend prediction. Time based sequence is prepared in this phase to provide input to LSTM neural network.

Phase 3: In this phase, LSTM Network is trained to perform prediction on time series data. LSTM network is capable to process long term patterns and able to determine how long to hold onto old information, when to remember and forget, and how to make connections between old memory with the new input. Here historical data is transformed to time series and existing data patterns are used to predict future trend pattern. Deep neural network is opted for technology classification from social text which is first component of model. After technology classification, quantification of data is performed

to map technology term with time parameter. Day-wise, demand of each technology is calculated in second phase of model development to generate temporal sequence for trend prediction. In second phase of model development, the researcher has created component to map technologies with time parameter. After classification and tagging of data, day wise technology demand count is calculated to analyse trend of technology over the years and prediction of technology trend can be done by learning from historical sequence of data. Following figure 2 presents the structure of data after mapping technology demands with timestamp.

Fig 3. Temporal Sequence Sample Data

Above generated sequence is passed as input to next model component LSTM for analyzing and prediction trend patterns from date-wise technology demands. For initial years, demand of python technology is less, which is showing increasing pattern gradually in data. After temporal sequence generation, trend prediction is performed in third phase. In last phase of model development process, trend predication component is developed to find up-down technology pattern from historical data and prediction of technology trend is done using deep LSTM network.

Researcher has developed model which can combine deep feed forward network and LSTM network to serve the purpose of technology classification and trend prediction both. This joint model implementation performs social media analytics and trend prediction task in single end-to-end architecture automatically.

For trend prediction from time sequence data, following model are experimented to evaluate performance of LSTM prediction.

IV. EXPERIMENT RESUTLS AND DISCUSSION

This section contains the details of experiment carried out using trend prediction models. These experiments are carried out on 10,950 timestamp records of programming technologies for year 2008 to 2014. It has been observed that LSTM model has performed better in context of time complexity and lowest error. Based on the results of experiment, LSTM model is selected for trend prediction component.

A. Technology Trend Prediction using Auto Arima

Auto regressive statistical models are applied for prediction of time series data and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an ARIMA algorithm builds a model that predict new data points based on analysis of past data trend pattern. Following configuration are selected for experiment and analysis. Auto ARIMA is applied with regression as predictor variable is separable linearly and having numeric value to predict. Model training details and model performance results are shown in Table 1 and Table 2.

TABLE I. AUTO ARIMA EXPERIMENTAL SETUP DETAILS

Auto Arima Parameters	
Sampling	Random
Cross Validation	3-fold
Train-Test Set	66%-34%

TABLE II. AUTO ARIMA MODEL EVALUATION

Model Training	MSE	RMSE	MAE	MPE
Training Set	0.6379	0.357	0.317	0.336
	0.6458	0.399	0.138	0.205
	0.6578	0.386	0.107	-0.167
Test Set	0.6110	0.367	0.090	0.145
	0.6277	0.379	0.095	0.152

B. Technology Trend Prediction using Linear Regression

Linear regression are supervised learning models with associated learning algorithms that analyze data used for and regression analysis. the regression model allows for a linear relationship between the forecast variable y and a single predictor variable x.Following configuration are selected for experiment and analysis. Linear Regression is applied with linear kernel as predictor variable is separable linearly and having numeric value to predict. Model training details and model accuracy results are shown in Table 4 and Table 5. The time complexity of linear regression model is 567s for training the dataset which is least compared to other models.

TABLE III. LINEAR REGRESSION EXPERIMENTAL SETUP DETAILS

Auto Arima Parameters	
Sampling	Random
Cross Validation	3-fold
Train-Test Set	66%-34%

TABLE VI. LINEAR REGRESSION MODEL EVALUATION

Model Training	MSE	RMSE	MAE	MPE
Training Set	0.5411	0.257	0.316	0.336
	0.5892	0299	0.138	0.205
	0.6578	0.386	0.107	-0.167
Test Set	0.6874	0.457	0.060	0.145
	0.5477	0.279	0.065	0.152

.

C. Technology Trend Prediction using LSTM

Support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. Following configuration are selected for experiment and analysis. SVM classifier is applied with linear kernel as predictor variable is separable linearly and having binary value to predict. Model training details and model accuracy results are shown in Table 2 and Table 3.

TABLE V. LSTM EXPERIMENTAL SETUP DETAILS

Auto Arima Parameters	
Sampling	Random
Cross Validation	3-fold
Train-Test Set	66%-34%

TABLE VI. LSTM MODEL EVALUATION

Model Training	MSE	RMSE	MAE	MPE
Training Set	0.6379	0.357	0.317	0.336
	0.6458	0.399	0.138	0.205
	0.6578	0.386	0.107	-0.167
Test Set	0.6110	0.367	0.090	0.145
	0.6277	0.379	0.095	0.152

V. CONCLUSION

Researcher has selected deep feed-forward neural network for technology classification as discussed in above sections and extended study of deep neural network architectures for making predictive model scalable to process data volume of year 2008 to 2016. As a part of methodology, this paper covers construction, refinement and development of Deep Feed Forward and LSTM based Trend Prediction model. Model construction started with technology trend prediction using hybrid neural network approach. Refinement of model is carried out by experiment with three different regression models for evaluation of proposed architecture. Temporal Sequence Mapping is created to quantify text data and generate sequence of date wise technology demand. Technology trend prediction component is made using LSTM network. Subsequent to deciding the deep neural network based approach, studies were extend to develop end-to-end generalized model where hybrid neural network architecture of Feed Forward Deep Network and LSTM was introduced. The research study has successfully integrated multitier data sources, managed data representation for model generalization to facilitate accurate data analytics and prediction for the future. This model has been validated by the use case of predicting the emerging technology trends of information technology domain.

VI. REFERNCES

[1] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, Wenbo Wang, Twitris: A System for Collective Social Intelligence, Encyclopedia of Social Network Analysis and Mining, ISBN 978-1-4614-6170- 8,2014J. Clerk

Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Mathieu Bastian, Matthew Hayes, William Vaughan, Sam Shah, Peter Skomoroch,Hyungjin Kim, Sal Uryasev, Christopher Lloyd, LinkedIn skills: largescale topic extraction and inference, Proceedings of the 8th ACM Conference on Recommender systems, ISBN 978-1-4503-2668-1 2014.

[3] S. Rill, D. Reinel, J. Scheidt and R. V.Zicarib. PoliTwi: Early detection of emerging political topics on twitter and the impact on concept-level sentiment analysis, Knowledge-Based Systems, Issue 1, Volume 69, pp. 24-33, 2014.

[4] Pasquale Lops, Marco de Gemmis, Giovanni Semeraro, Fedelucio Narducci, Cataldo Musto, Leveraging the linkedin social network data for extracting content-based user profiles, Proceedings of the fifth ACM conference on Recommender systems, ISBN 978-1-4503-0683-6 2011.

[5] Mikhail Klassen, Matthew A. Russell. Mining the Social Web, 2nd Edition, O'Reilly

[6] Dean Abbott, Applied Predictive Analytics, Indianapolis: Willey India, 2014.

[7] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang. Traffic Flow Prediction With Big Data: A Deep Learning Approach, IEEE Transaction on Intelligent Transport System, Issue 2, Volume 16, pp. 865-873, 2015.

[8] Sacide Güzin Mazman, Yasemin Koçak Usluel. Modeling educational usage of Facebook, Computers & Education, Issue 2, Volume 55, pp. 444-453, 2010

[9] A. Begel, J. Bosch, and M.A. Storey, "Social networking meets software development: Perspectives from github, msdn, stack exchange, and topcoder," IEEE Software, Issue 1, Volume 30, pp. 52 - 66, 2013

[10] A. S. Rathor, A. Agarwal and P. Dimri, "Comparative Study of Machine Learning Approaches for Amazon Reviews," in International Conference on Computational Intelligence and Data Science, 2018.

[11] L. Le, E. Ferrara and A. Flammini, "On predictability of rare events leveraging social media: a machine learning perspective," in Proceedings of the 2015 ACM on Conference on Online Social Networks, California, 2015.

[12] J. L. Hurtado, A. Agarwal and X. Zhu, "Topic discovery and future trend forecasting for texts," Journal of Big Data, vol. 3, no. 1, 2016

[13] Chen, Q. Kong, N. Xu and W. Mao, "NPP: A neural popularity prediction model for social media content," Neurocomputing, vol. 333, p. 221–230, 2019.

[14] Z. Zhang, Q. He, J. Gao and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," Transportation Research Part C: Emerging Technologies, vol. 86, pp. 580-596, 2018.

[15] H.-B. K. Hyeon-Woo Kang, "Prediction of crime occurrence from multi-modal data using deep learning," PLOS ONE, vol. 12, no. 4, 2017.

[16] Q. Zhang, L. T. Yang and Z. Chen, "Deep Computation Model for Unsupervised Feature Learning on Big Data," IEEE Transactions on Services Computing , vol. 9, no. 1, pp. 161-171, 2016.

[17] R. G. Guimaraes, R. Rosa, D. D. Gaetano and D. Z. Rodriguez, "Age Groups Classification in Social Network Using Deep Learning," IEEE Access, vol. 5, p. 10805–10816, 2017.

[18] S. J. M. I. H. S. P. M. Dat Tien Nguyen, "Applications of Online Deep Learning for Crisis Response," in Springer, 2016.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," Computation and Language, vol. 3, 2013.

[20] S. Desai and S. Patil, "Efficient regression algorithms for classification of social media data," in International Conference on Pervasive Computing (ICPC), Pune, 2015.