Deep Learning Model Development to Classify Technology from Social Media

Yesha Mehta[1]**, Kalpesh Lad**[2]

[1]Department of Computer Science, Shree Ramkrishna Institute of Computer Education and Applied Sciences, Surat, India.
[2]Department of Computer Science, Shrimad Rajchandra Institute of Management and Computer Application, Bardoli, India.

Publication Info

Article history:
Received : 15 February 2020
Accepted : 21 May 2020

Keywords:
Deep learning, Technology classification, Social media, Deep neural network.

Corresponding author:
Yesha Mehta
e-mail: *yesha.mehta@srki.ac.in*

Abstract

The important characteristic of a societal platform is the unremitting cohort of content, which leads to the root of novel data out of it. Community platforms are active in nature, and they can be painstaking as a new type of data resource for upcoming trend forecasts with the application of data analytics techniques. Due to the world-wide contribution to open-source technologies, new technologies are released frequently. The innovation and instruments learned by individuals become supplanted in a brief timeframe. As the I.T. industry required regular overhauls in abilities and new advancements are being discharged, it is basic to trail and realize novel upcoming innovation patterns of the field. To accomplish this point, a deep learning model is created to distinguish up and coming innovations from internet-based life strings. This paper presents a technology classifier model to categorize unstructured text content into relevant technologies. To develop technology classifiers, classification algorithms Support Vector Machine (SVM), Decision Tree, k Nearest Neighbor (kNN), Artificial Neural Network (ANN), and Deep Feed-Forward Neural Network are trained and experimented with forecasting expertise terms from unstructured content of social posts. The results of the experiment show that Feed-Forward deep neural network outperforms other classification models and provides better accuracy and technology term prediction.

1. INTRODUCTION

The I.T. industry alterations are frequent, and due to that, people engaged with the I.T. profession have to keep their self-updated with technology expansions. I.T. professionals and students strive to learn new technologies to prepare themselves for their next challenging I.T. role. Due to worldwide contribution to open-source technologies, new technologies are released regularly. Technology changes at a speed of light, so working in the I.T industry creates a need to be updated about the state-of-the-art technologies. Trend analysis is generally performed on quantitative data. To analyze trends, form data posted on a social platform is a challenging task because of the unstructured format of content, and lack of quantities from data. The researcher has studied approaches where social media platforms Twitter and LinkedIn are analyzed for sentiment analysis, citizen sensing [1], prediction of city events [2], skill topic extraction [3], and user profile analysis [4]. The key trait of the social stage is the ceaseless age of substance, which prompts the inference of new information out of it. Social stages are dynamic in nature, and they can be considered as new kinds of data assets for future pattern expectations with the utilization of information examination procedures. The conversation strings and substance accessible on the sites as Twitter, Facebook, and other online interfaces remembers most recent data and individuals' view for various subjects and themes.

Social Media data is different from traditional documents such as newspaper articles. These non-structured texts can be found in several formats, written by dissimilar people in numerous languages and styles, written in day-to-day language. It is a scientific challenge to develop an algorithm which extracts relevant information or topic from the text.

Utilization of unstructured social substance for tackling understand up and coming innovation forecast sort of issue is trying because of the intricacy of the structure, absence of setting, and casual nature of language. Considering the part of critical thinking, the substance assorted variety and information volume intrinsic in web-based life make huge down to earth difficulties for removing important data. Recognizing the "Rising Technology Trends" from internet-based life is an open test and engaging work that prompts the investigation of introduced look into the issue.

This kind of research work can help partners of instruction network, including understudies, scholarly foundations, the executives, college, and so forth to help their dynamic procedure in zones of business, vocation way choice, and educational program planning process. To handle the problem of social media mining and analysis, novel data mining techniques are required to be developed to analyze content created by multiple and heterogeneous users. In traditional data mining systems, the information is in a highly structured format and having a fixed volume compared to today's data sources. Existing data mining tools can not directly incorporate unstructured data formats to identify or extract knowledge [9]. The motivation and idea behind this research are to design and develop a concept for processing open social media data and derive knowledge about emerging technologies of the I.T. field.

2. TECHNOLOGY CLASSIFICATION MODEL DEVELOPMENT

To accomplish the objective of technology trend prediction, there is a need to develop a process that can continuously collect and analyses information from social media. Developing a model for technology topic classification is challenging due to the large amount as well as the complexity of data formats. In this section, the model development process is explained where unstructured data is processed, and a classifier is developed to categorize unstructured text content into relevant technology topics.

2.1. Data Sources

Emerging technologies are specified in the job requirements which are posted by companies on job portals. There are online platforms and open forums to discuss technology-related queries. Social platforms capture instant updates in discussion threads regarding any upcoming technology and trends. Considering all potential sources which capture the technology updates in their data itself, the following data sources are selected as information resources for the presented model. To select appropriate data sources, four types of data sources are analyzed, which include real-time data from users, social discussion content, and technology demand in job requirements.

- Social media - Twitter
- Question answer sites - Stackoverflow
- Job portals - Monster, Indeed
- I.T. project site - GitHub

These four data sources are selected to perform experiments for developing a technology classification model. The underlying dataset is having social post and their technology label for classification purposes. Input data of the model will have textual content with technical terms, and the developed model is aimed to identify and predict technology label/tag social post. To achieve this aim, classification models are experimented to predict technology from unstructured content of social posts.

2.2. Classification Algorithm Selection

The researcher has performed a comparative analysis of classification algorithms to study the characteristics, strengths, and limitations for selecting a classification algorithm to develop a technology trend classifier. Table 1 presents the comparative study of classification algorithms in the context of their implementation method. The classification model is selected with consideration of the type of input data for this research work, the volume of the data, computational complexity, and time complexity. It has been derived from a comparative study and literature survey that SVM, Decision Tree, KNN, and ANN algorithms are compatible with data sources of social platforms.

3. PRE-PROCESSING SOCIAL CONTENT

The data sources Twitter, Stackoverflow, and job portals have unstructured text content with the following characteristics.

- Social posts and content are created by multiple users, i.e., job post or instant messages.
- The unstructured text does not have information on metadata, and it is difficult to map this information to standard database fields.

Table 1: Comparative study of classification algorithms

Classification algorithms	*Characteristics*
Naive Bayes	This classification method implements Bayes' Theorem of probability, which implicitly assumes that all the attributes of data are mutually independent.
Support vector machines	Machine learning algorithm which divides the data in n-dimensional planes for classification purpose.
Decision trees	This algorithm is applied with decision rules and a tree structure to perform predictions from categorical and numerical data.
Artificial neural network	This algorithm implements weights and biases in the linear equation and calculates the probability of an outcome based on calculations of mathematical units called neurons.
K-nearest neighbor	Algorithm which works on Euclidean distance between data points and classify new data points based on nearest observations.
Deep neural network	Learns underlying representations from data itself by deep neural layers and classify observations based on historical data.

- Due to a lack of identifiable structure, it cannot be used by computer programs without pre-processing.
- Data neither conforms to a data model nor has any structure.

Unstructured text data found in informal content of social platforms and web pages are noisy and dynamic; there is a need to process it for the information retrieval [17]. Data pre-processing transforms social content for further processing. This component consists of sub-processes data cleaning, tokenization, and transformation.

3.1. Data Cleaning

As social post contains misspelled words, quotations, program codes, extra spaces, extra line breaks, special characters, foreign words, etc. thus in order to achieve high-quality text content, it is necessary to conduct data cleaning at the first step.

The following steps are performed for text cleaning.

- HTML decoding,
- Changing text to lower case,
- Removal of junk characters,
- Removal of stop words,
- Removing spaces, punctuations, and additional junk characters from the text.

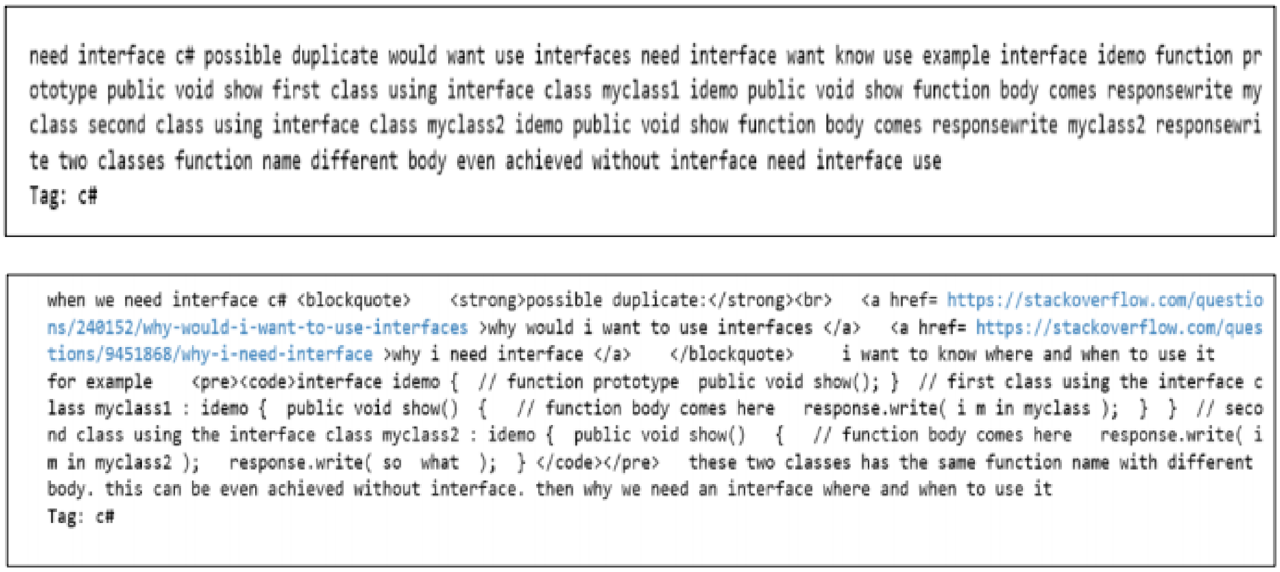
3.1.1. Algorithm 1: An algorithm for social post text content cleaning

```
STEP 1.START
STEP 2.Define special symbols and characters to be replaced in text.
        Set REPLACE_BY_SPACE_RE =re.compile('[/(){}\[\]\|@,;]')
STEP 3.Define junk characters
        JUNK_SYMBOLS_RE = re.compile('[^0-9a-z #+_]')
STEP 4.Define stop words
        STOPWORDS = set(stopwords.words('english'))
STEP 5.Read job_post , tweet, question texts from of csv files
STEP 6.FOR each post until length
        a. HTML decoding
           text = BeautifulSoup(text, "lxml").text
        b. Lowercase text text.lower()
        c. REPLACE_BY_SPACE_RE symbols by space
           REPLACE_BY_SPACE_RE.sub(' ', text)
        d. Delete symbols
           JUNK_SYMBOLS_RE.sub('', text)
        e. Delete stopwords from text
           ' '.join(word for word in text.split() if word not in STOPWORDS)
STEP 6.END FOR
STEP 7. END
```

Above explained algorithm is developed by researcher to clean the unstructured textual content of social posts which will clean the text data for further processing. The sample of extracted post is presented where difference in content before text data cleaning and after cleaning can be identified from sample given in Figure 1.

3.2. Tokenization and Transformation

To separate and filter technology terms from the social post shown in Figure 1, the process of tokenization is performed. In tokenization, there is a requirement to filter tokens out of social posts. Tokenization helps to divide the



```
need interface c# possible duplicate would want use interfaces need interface want know use example interface idemo function pr
ototype public void show first class using interface class myclass1 idemo public void show function body comes responsewrite my
class second class using interface class myclass2 idemo public void show function body comes responsewrite myclass2 responsewri
te two classes function name different body even achieved without interface need interface use
Tag: c#
```

```
when we need interface c# <blockquote>   <strong>possible duplicate:</strong><br>   <a href= https://stackoverflow.com/questio
ns/240152/why-would-i-want-to-use-interfaces >why would i want to use interfaces </a>   <a href= https://stackoverflow.com/ques
tions/9451868/why-i-need-interface >why i need interface </a>   </blockquote>   i want to know where and when to use it
for example   <pre><code>interface idemo {  // function prototype  public void show(); }  // first class using the interface c
lass myclass1 : idemo {  public void show()  {   // function body comes here   response.write( i m in myclass );  }  }  // seco
nd class using the interface class myclass2 : idemo {  public void show()   {   // function body comes here   response.write( i
m in myclass2 );   response.write( so  what  );  } </code></pre>   these two classes has the same function name with different
body. this can be even achieved without interface. then why we need an interface where and when to use it
Tag: c#
```

Figure 1: Stackoverflow post sample – before and after data cleaning

textual information of social posts into individual words. Each sentence is divided into words using a tokenizer. In order to pass the words into the neural network, the tokens are required to be transformed into vector representations. There is an encoding scheme known as one hot vector encoding, which is a representation of categorical variables as binary vectors. Tokens are sorted and assigned an integer value. The occurrence of the term in a sentence is represented as a binary vector that is all zero values except the index of the integer, which is marked with 1. Social text is converted in tokens. The set of tokens is arranged in alphabetical order, and their occurrence in the text is noted by a pattern of integer encoding. Integer encoding is converted into two-dimensional metrics of 1 and 0.

Sample text: Python for NLP: Developing an Automatic Text Filter N-Grams

Cleaned text: Python NLP: Developing Automatic Text Filter N-Grams

Word positions: 0 1 2 3 4 5 6

Tokens: [‘python’, ‘nlp’, ‘developing’, ‘automatic’, ‘text’, ‘filter’, ‘n-grams’]

Alphabetical order: [‘automatic’, ‘developing’, ‘filter’, ‘n-grams’, ‘nlp’, ‘python’, ‘text’]

Integer encoded (position of occurrence of word in sentence): [3, 2, 5, 6, 1, 0, 4]

Matrix representation:

```
[[ 0 0 0 1 0 0 0],
 [ 0 1 0 0 0 0 0],
 [ 0 0 0 0 0 1 0],
 [ 0 0 0 0 0 0 1],
 [ 1 0 0 0 0 0 0],
 [ 0 0 0 0 1 0 0]]
```

4. TECHNOLOGY CLASSIFIER

After the pre-processing of social posts content, SVM, Decision Tree, kNN, and Deep Neural Network algorithms have experimented on data sources described in Section 2. Based on experiments, algorithm selection for technology classification is performed by observing the classification accuracy of the model.

4.1. Technology Classification using SVM

Support-vector machines are first applied to classify social posts in and predict their relevant technology tag. The classification model is trained with 66% of training and 34% of testing data on the dataset of 1,50,000 social posts extracted from heterogeneous sources. Given a set of training examples, each marked as belonging to one or the other of two categories, and an SVM training algorithm builds a model that assigns new examples to one category or the other. SVM classifier is applied with the linear kernel as the predictor variable is separable linearly and having binary value to predict. The model accuracy results are shown in Table 2.

4.2. Technology Classification using kNN

After SVM, an experiment is carried out with the kNN algorithm, which is a non-parametric method used for classification and regression. The kNN model is trained by selecting Euclidian distance, random sampling, and 3-fold cross-validation. The model is evaluated on 34% of the test set, where the following results are achieved for classification. It has been observed that the kNN model is suitable with seven neighbors and 89.40% of result accuracy.

4.3. Technology Classification using Deep Neural Network

A deep neural network model is selected by the researcher with the intent to execute experiments with large data volume and development of a model with the skill to mechanically study optimum feature representation for

Table 2: SVM classifier model evaluation

Iterations	*Accuracy*	*Precision*	*Recall*	*F1-score*
50	84.00	0.357	0.317	0.336
80	85.80	0.399	0.138	0.205
100	85.90	0.386	0.107	0.167
120	86.70	0.367	0.090	0.145
140	86.70	0.379	0.095	0.152

Table 3: K-N classifier model evaluation

Neighbours K	*Accuracy*	*Precision*	*Recall*	*F1-score*
2	84.70	0.357	0.317	0.336
3	88.80	0.399	0.138	0.205
5	88.50	0.386	0.107	0.167
7	89.40	0.693	0.278	0.145
8	89.20	0.379	0.095	0.152

Table 4: Deep neural network classifier model evaluation

Neurons	*Accuracy*	*Precision*	*Recall*	*F1-score*
100	90.30	0.125	0.125	0.222
150	91.70	0.199	0.138	0.205
200	93.10	0.186	0.107	0.245
250	94.40	0.167	0.090	0.208
300	94.40	0.179	0.095	0.208

technology classification. With a huge measure of preparing information, profound neural systems can proficiently delineate the crude contribution of content to a low-dimensional vector portrayal, which jelly syntactic just as semantic parts of the info content. The following setup is chosen for trial and examination. DNN classifier is applied with ReLu actuation as an indicator variable is divisible straightly and having double an incentive to foresee. Model preparing subtleties and model exactness results have appeared in Table-4.

Model optimization is performed by a training model with hidden layer neurons of 50, 100, 150, 200, and 250. It has been observed that there is no improvement in classification accuracy after 200 neurons in the hidden layer, where the threshold limit of 200 neurons is accepted for validating the model on the test dataset, which gave 94.40% of classification accuracy. SVM, Decision Tree, kNN, ANN, and Deep Feed-Forward algorithms are selected for the experiment [8]. The accuracy rate obtained in SVM, kNN, Decision Tree, and dense feed-forward network is 86.70, 89.40, 99.63, and 94.40%, respectively. Based on the experiment result, a deep feed-forward neural network is selected for technology classification.

5. DEEP FEED FORWARD NEURAL NETWORK ARCHITECTURE

The researcher has designated a deep feed-forward neural network for technology classification as debated in the above section and protracted study of deep neural network architectures for making predictive model scalable to process data volume of the year 2008 to 2016. Firstly, basic experiments are done to predict trends of three main technologies (java, .net, python) on 14,10,198 records. Based on the experiments, the deep neural network is chosen for technology classification from community writing. The following figure presents the process flow of deep neural network architecture to classify and tag social posts into technology topics.

As shown in Figure 2, social post data is extracted first and pre-processed before passing to the predictive model. A deep neural network is trained with labeled data of social posts and its relevant technology tags. The trained model is tested on the unseen test data where on social post content is given to model in the form of vectors, and technology tag is predicted by the model with an accuracy of 94.4%, which is a good benchmark for heterogeneous data sources. The data passed as an input model is converted into tokens and word-vectors to be processed by a dense layer of the model. After a dense layer, the ReLu activation layer is added to remove the effect of negative values, and again Dense Layer is added for training the model. At last Softmax layer is added to calculate the probabilities of the class label where tag having the highest probability is selected to classify the post.

As a portion of the procedure, this paper covers the structure, modification, and growth of the Technology Classifier model. Feed-Forward deep neural network is selected for technology classification with the highest accuracy of 94.40%. After the development of the Technology Classifier, the Temporal Sequence Mapper component is planned to enumerate text information and produce an order of date wise technology request. Following determining the deep neural network-based method, studies were spreading to develop an endwise generalized model where hybrid neural network architecture of the feed-

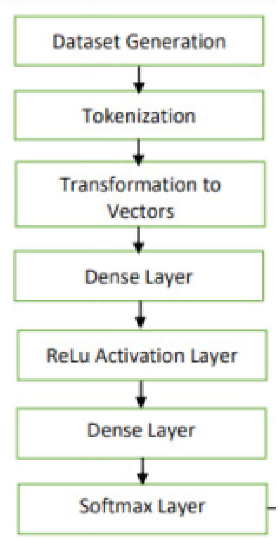


Figure 2: Deep neural network technology classification architecture

forward deep network and LSTM is premeditated for trend forecast purposes. The designed model is developed as part of the social media analytics framework.

6. CONCLUSION

This paper presents a study of the deep learning model development for the classification of social posts into technology topics. Based on the experiments, deep neural network model archives highest accuracy compared to other classification algorithms and selected for making the technology classifier model. Pre-processing discusses the data extraction, data pre-processing techniques utilized in this study, and data transformation techniques. This component uses pre-processing techniques of text cleaning, tokenization, and vector representation. The approach of focusing upon automatic identification of technology terms and the maximum likelihood of prediction accuracy has resulted in deep learning hybrid neural network models as the most suitable model over machine learning algorithms and statistical models. Results achieved as a part of our research experiment related to the prediction of future trends and technologies are aligned and matching with job demands. The model developed during our research work is able to extract the data from social platforms and identify the technology topics of the I.T. field and tag them without a manual feature engineering process. This work can be extended further to eradicate restriction of the feed-forward deep neural network as it is not designed for time-dependent data. The problem of technology trend analysis and prediction needs to process data, which is consuming time parameters or time-based order patterns where the possibility of future improvement is there.

7. REFERENCES

7.1. Journals

[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.

7.2. Books

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

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

7.3. Conference Proceeding

[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.