International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Topic Detection using Machine Learning

Mr. Ajmal Rasi, Dr. Rajasimha A Makram, Ms. Shilpa Das
School of Engineering and Technology Jain University, Bengaluru, Karnataka, India

ABSTRACT
Various types of social media such as blogs, discussion forums and peer-to-peer networks present a wealth of information that can be very helpful. Given vast amount of data, one of the challenge has been to automatically identify the topic of the background chatter. Such emerging topics can be identified by the appearance of multiple posts on a unique subject matter, which is distinct from previous online discourse. We address the problem of identifying topics through the use of machine learning. I propose a topic detection method based on supervised machine learning model, where sentences are labelled, tokenized and the vectorised sentence is trained on densely connected neural network. Compared to conventional gradient descent optimization algorithm, Adam optimizer trains the data much faster and efficiently. Finally the model is tested on an Android App with live data from Google News.

Keywords: Machine Learning, Supervised Learning, Neural Networks, Topic Detection, Natural Language Processing

1. INTRODUCTION

With the explosion of web, various types of social media such as blogs, discussion forums and peer-to-peer networks present a wealth of information that can be very helpful. Driven by the demand of gleaning insights into such great amounts user-generated data, work on new methodologies for automated text classification and discovering the hidden knowledge from unstructured text data has bloomed splendidly. Given the vast volume of such data being continually generated, one of the challenges is to automatically tease apart the emerging topics of discussion from the constant background chatter.

Artificial intelligence learning generally consists of two main methods. These include supervised and unsupervised learning. Supervised learning involves using an existing training set, which consists of pre-labelled classified rows of data. The machine learning algorithm finds associations between the features in the data and the label for that row. In this manner, a machine learning model can predict on new rows of data, that it has never been exposed to prior, and return an accurate classification based upon its training data. Supervised learning can work great for large data sets where you already have pre-classified data readily available.

2. The Proposed System.

The wide-spread popularity of social media, such as blogs and Twitter, has made it the focal point of online discussion and breaking news. Given the speed at which such user-generated content is produced, news flashes often occur on social media before they appear in traditional media outlets. Twitter, in particular, has been at the forefront of updates on disasters, such as earthquakes, on the post-election protest, and even on news of celebrity deaths. Identifying such trending topics is of great interest beyond just reporting news, with applications to marketing, disease control, national security and many more.

- Identify topic form a sentence using supervised machine learning model where high performance systems are not available for unsupervised model
- Create the algorithm using Python on PC
- Train the model with neural network algorithm
- Save the model and export it to Android environment
- Inference the model and predict results on live data

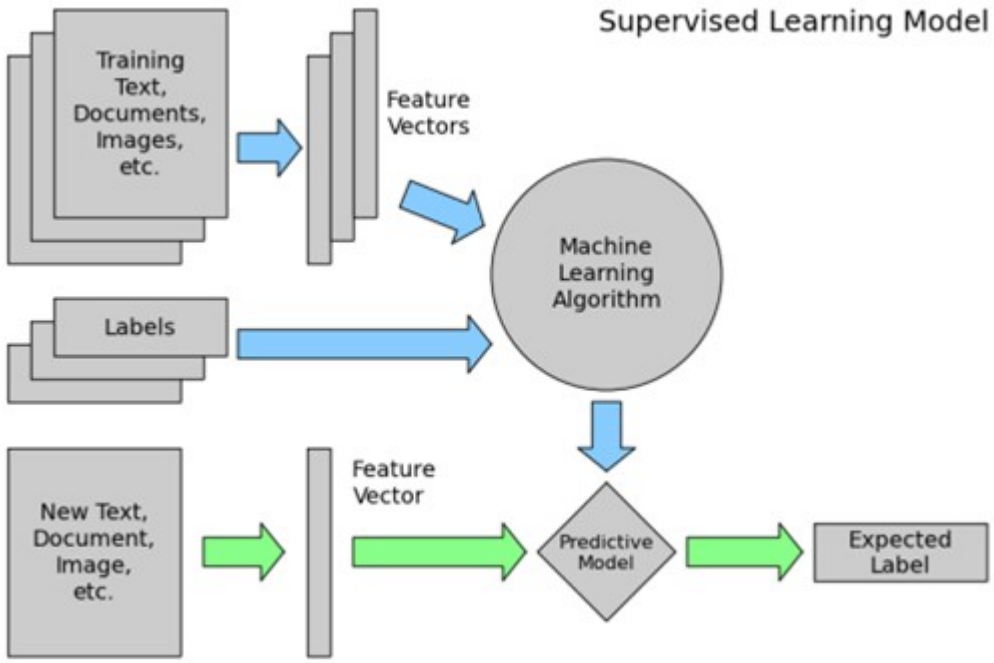

Figure 2. 1 Proposed System Architecture

3. Implementation

3.1 Data Acquisition

Data acquisition has been understood as the process of gathering and filtering of data before the data is put in a storage solution. Text data here are news titles from RSS feeds. Now the first step is to collect the kind of data you would need for your analysis.

Google provide news in XML, a page has 20 <item> tag. Every <item> tag contains link, title, category and description of news articles. But for training purposes we need only the title. To do that this data has to be parsed. First data is read into a remote web server using curl_ method in PHP, then this retrieved RSS data is parsed with SimpleXMLElement. By passing the key as title, title text is queried. This process is done for every item in the RSS data, then this queried text is saved as a CSV file. But the above method fetches only handful of data, 20 items to be precise. To capture more the above process has to be repeated many times during the data

Figure 3. 1 XML News Data

Instead of doing this manually, CPanel offers a functionality called CRON Jobs, which is automatic execution of given script at certain intervals throughout the day. Now this is applied for the above program, which runs 96 times a day or runs every 15 minutes. This process produces a lot of data, all of which is saved in remote server.

3.2 Data Labelling

Before labelling data, CRON jobs produces 96 JSON files every day, all the titles inside 96 JSON files is appended to a single JSON file with date as the name (e.g. 21-05-2018.json). Now this is loaded to a webpage containing input area, where the labels are given. This is repeated for every title in that, it is a tedious process, takes a lot a time and are prone to human errors. This is the main disadvantage of this process. After labelling, this file is again saved as JSON, now this time with labels. Data required for training is now ready to be downloaded for pre-processing

3.2 Data Preprocessing

Real-world data are noisy, missing and inconsistent, so after collecting the data, the data should be cleaned before any further processing. Cleaning textual data involves fixing typos, stemming and removing extra symbols and stop words. Another important task in data preparation is formatting. It is important that the collected data is in correct format. The data fed into model needs to be in correct format. Also, it is tempting to use all the data that is available. Larger dataset does not always guarantee high performance. In fact larger the dataset higher the computational cost. So, it is better to use a subset of the available data in the first run. If the smaller subset of the data does not perform well in terms of precision and recall, there is always an option to use the whole dataset. Some other common tasks that are done while preparing text data include tokenization, part-of-speech tagging, chunking, grouping and negation handling. Each of these tasks will be discussed in separate blog posts.

3.3 Training

Network is trained on input data in vector format of n dimensional matrix. n is the number of words (bags of words) present after pre-processing. There are two hidden layers with linear activation and the output layer is Softmax. Hidden layers having 8 nodes (Neurons) each and output is a m dimensional array with its probability, m is the number of topic.

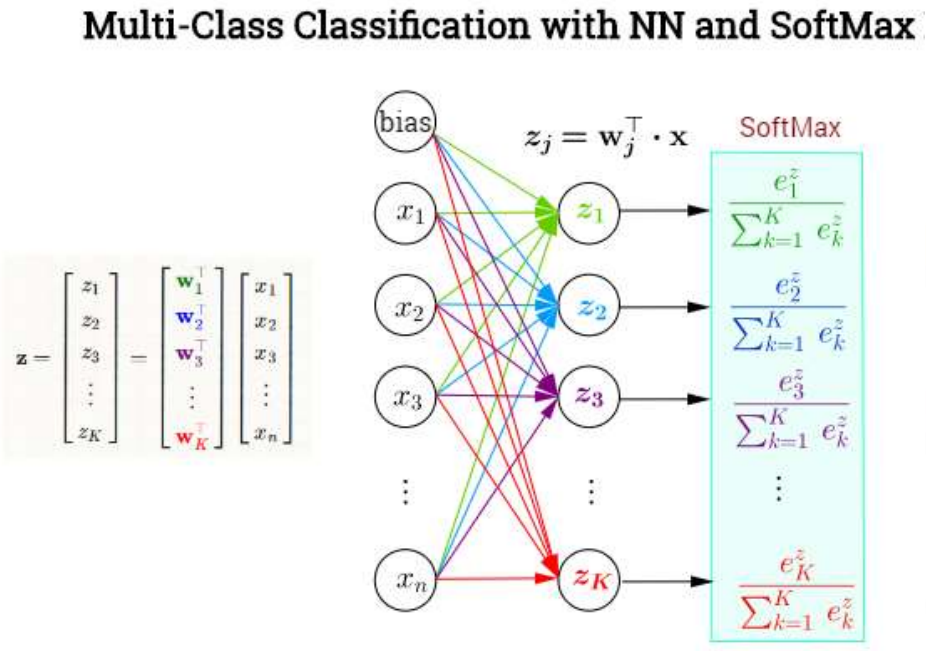

Figure 3. 2 Model Diagram

3.4 Testing

Testing is done after pre-processing titles and fed into the model as vectors. Output from the softmax layer is the probability distribution in m dimensional array. To find the label with highest probability argmax is used, which returns index of array with highest probability. Finally the results are predicted on test data.

3.5 Deployment

Up until now all the training and testing is done on a PC, our goal is to make this model work on android app. Save serialized model is imported to android studio as a Protocol Buffers file. Using Tensor Flow in ferencing library. Prediction are made, live data from News is fed to the model after pre-processing like we did earlier, but android is developed on java, there needs to be a java library which does pre-processing. The libraries used are Open NLP and Stanford NLP. Open NLP for stemming and tokenization and Stanford NLP for argmax function

4. Conclusion

The wide-spread popularity of social media has made it the focal point of online discussion and breaking news. Given the speed at which such user-generated content is produced, news flashes often occur on social media before they appear in traditional media outlets. Twitter, in particular, has been at the forefront of updates on disasters, such as earthquakes, on the post-election protest, and even on news of celebrity deaths. Identifying such trending topics is of great interest beyond just reporting news, with applications to marketing, disease control, national security and many more. with this method it is guaranteed to find accurate topic given a good enough data set during training.

5. REFERENCES

1) Jason Brownlee, Gentle Introduction to the Adam Optimization Algorithm for Deep Learning, 1-5, July 2017

2) Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, Reuven Y. Rubinstein, A Tutorial on the Cross-Entropy Method. 95-104, 2004 Dec 18

3) Alberto Quesada, 5 algorithms to train a neural network, 15 Feb 2017.

4) Christian Wartena, Rogier Brussee, Topic Detection by Clustering Keywords, 6 Apr 2007

5) Chenghua Lin., Yulan He, Stefan Rüger. Weakly Supervised Joint Sentiment-Topic Detection from Text. N. Engl. J. Med. 2015; 372:793–795

6) Diederik P. Kingma, Jimmy Ba, Adam: A Method for Stochastic Optimization, 2014

7) Tomáš Mikolov, Martin Karafiát, Lukáš Burget, The Personalised medicine. Recurrent Neural Network Based Language Model, 2010