DOMAIN IDENTIFICATION

A PREPRINT

Vivek Anand
IRE Lab, IIIT-H
vivek.a@research.iiit.ac.in

Yash Agrawal
IRE Lab, IIIT-H
yash.agrawal@research.iiit.ac.in

April 28, 2019

ABSTRACT

The project deals with identifying the domain of given text (specifically news articles). Domain identification is useful in case of machine translation/summarization tasks because if the domain is identified, then the translation/summarization models can be supplemented with domain knowledge. So the primary goal is to build a model that given a news article determine the domain from which the article belongs.

The project is extended to include two other parts as well. First is, once the domain of the news article is determined we would like to extract out the domain specific keywords from the document as well. Second, apply the same model to news articles in a language other than on which the model was trained (English here) and observe the results.

1 Approach

For the domain classification task we explored classical ML techniques and Deep learning techniques.

First, pre-processing of the news articles was done which included: case folding, tokenization, removal of stop words, punctuation marks, padding and/or trimming to a fixed length.

1.1 Classical ML methods for classification

- 50 Dimension pre-trained GloVe embedding was used for each token.
- We take the average of all the word embeddings and get a 50 Dimensional representation for the whole news article.
- This representation vector was then passed through SVM and Logistic Regression models and they were trained.

1.2 Deep Learning method for classification

- We trained bi-directional LSTM on vector representation of pre-processed news articles.
- In this method, rather than using pre-trained embeddings, word embeddings were also trained jointly.
- The last hidden state of LSTM is taken and passed through a single layer feed forward neural network to output probability values for the classes.

1.3 Keyword extraction from the article

- An attention layer is added on top of Bi-LSTM model for the domain classification task described above.[1].
- A random seed was initialized which was trained for generating attention weights using the hidden states of the Bi-LSTM.
- We then extract the keyword based on which words are attended the highest for the classification of the domain.

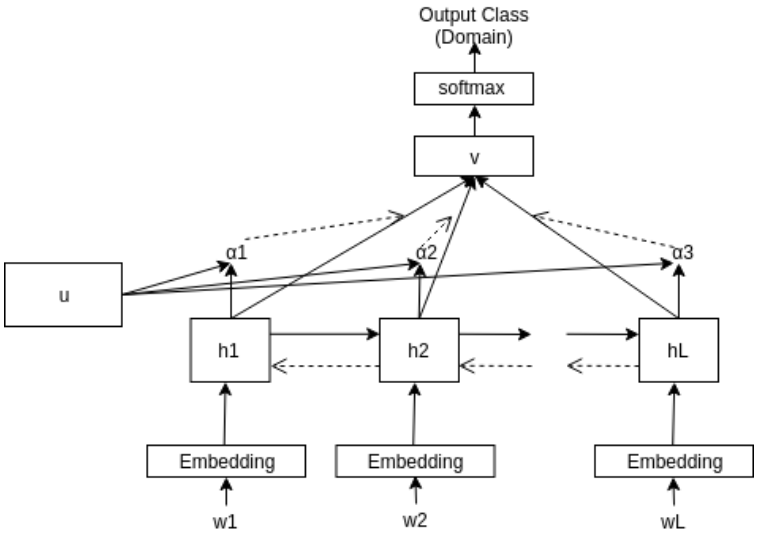

Figure 1: Attention Based Classifier.

- The idea behind this approach is that words more specific to the domain will be attended more in classification of the document for a particular domain.

2 Dataset

16705 news articles across the domains are scrapped from news websites- [www.hindustantimes.com, www.timesnownews.com, www.livemint.com, www.theguardian.com, www.techcrunch.com, www.crictracker.com, www.phys.org].

These news articles belong to 7 domains: ['business', 'science', 'politics', 'entertainment', 'sports', 'technology', 'automobile']

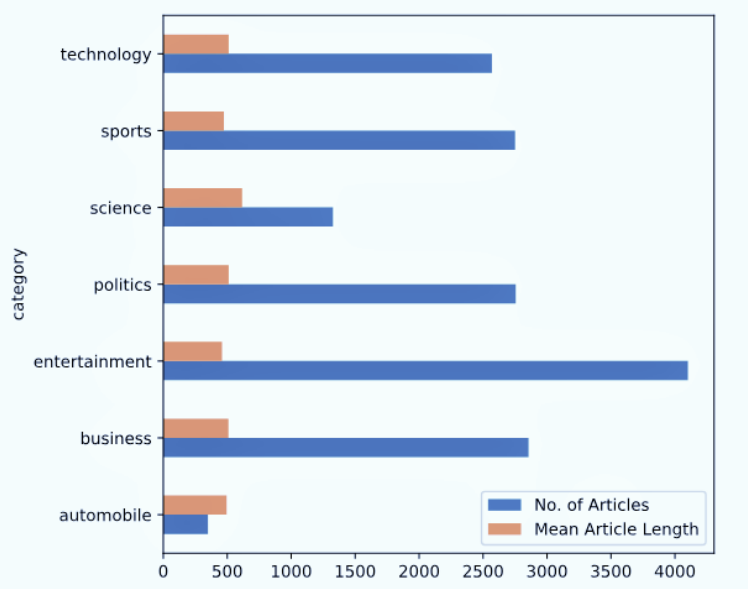

Figure 2: Articles Domains and Mean Length (in words)

3 Experimental Setting

- We used Scikit learn for SVM and logistic regresison implementation.
- Keras was used for the implementation of Bi-LSTM classification model.
- We used Pytorch as the deep learning framework for attention based keyword extraction and classification model.
- Stochastic gradient descent with batch size of only 1 was used to minimize the loss.
- The best model we got was at 11th epoch and after that the model started over fitting.
- For training of deep learning models, GTX 1080 Nvidia GPU was used which is courtesy of IRE Lab.

4 Results

4.1 Classification of Domain

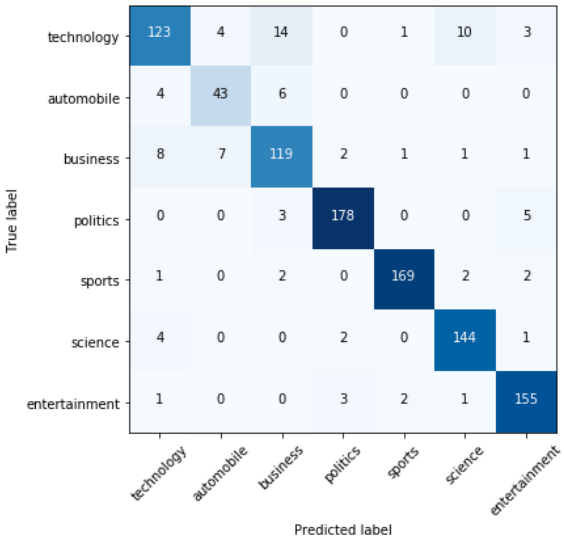

Figure 3: SVM Classifier

Table 1: Accuracy of different classifier

Classifier	Accuracy
Logistic Regression	90.01 %
SVM	91.09 %
Attention + Bi-LSTM	91.58 %

4.2 Keyword Extraction

References

[1] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola and Eduard Hovy. Hierarchical Attention Networks for Document Classification.

Table 2: Some examples of extracted keywords

Picking specialists is an extension of Virat Kohli's Test selection policy but this time it has produced a World Cup squad that looks really thin on batting. Going by the primary role of the players, the 15-member India squad has five bowlers — Jasprit Bumrah, Mohammed Shami, Bhuvneshwar Kumar, Yuzvendra Chahal and Kuldeep Yadav – four all-rounders — Kedar Jadhav....
Domain- Sports
Keywords- ['left-handed', 'legendary', 'sourav', 'gautam', 'ganguly', 'gambhir', 'cups', 'batsman', 'tournament', 'best']
SBI Capital Markets is said to have shortlisted four prospective bidders for the stake sale in Jet Airways as a part of the debt resolution process. These include TPG Capital, Indigo Partners, Etihad Airways and NIIF. The qualified bidders are expected to submit their binding bids latest by April 30. It is a crucial process for Jet Airways that is battling a number of issues on multiple fronts to stay operational. Here's a quick look at the bidders that are....
Domain- Business
Keywords- ['shortlisted', 'prospective', 'sale', 'four', 'stake', 'markets', 'carrier', 'said', 'debt', 'airline']
If you have read my Redmi Note 7 Pro review (read here) you'll know how much I loved the stunning design of the phone. I feel the same about the Redmi Note 7. I'm absolutely in love with the design of the Redmi Note 7, thanks to its all-glass and gradient finish, minimal bezels, and dot drop notch. The design of the Note 7 is....
Domain- Technology
Keywords- ['smartphones', 'phones', 'build', 'cheaper', 'segment', 'price', 'body', 'best-looking', 'slightly', 'definitely']
NCP spokesperson Nawab Malik's remarks came after the Maharashtra BJP sought to know which Lok Sabha candidate would incur expenses of the poll rallies of Thackeray, whose party has not fielded any candidate for the Lok Sabha polls. Thackeray....
Domain- Politics
Keywords- ['thackeray', 'sabha', 'bjp', 'lok', 'mps', 'candidate', 'mns', '2014', 'leader', 'supported']

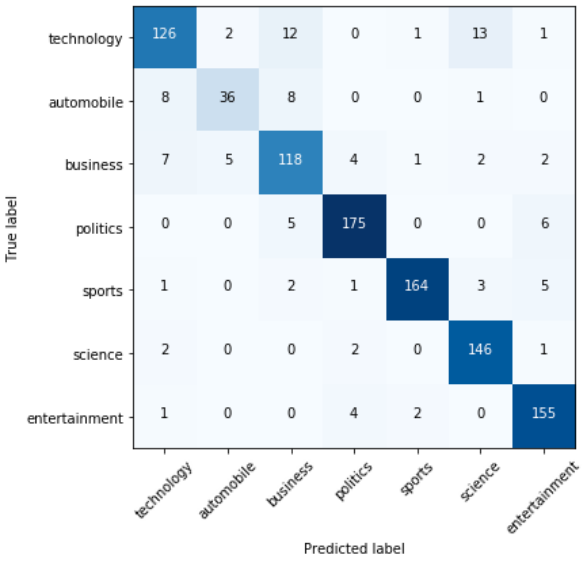

Figure 4: Logistic Regression Classifier

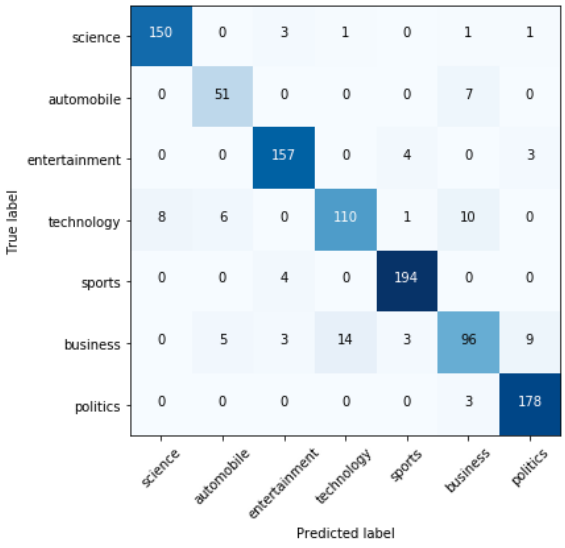

Figure 5: Attention Based Bi-LSTM Classifier