Article title: News Sentiment Analysis By Using Deep Learning Framework

Authors: Sweta Kaman[1]

Affiliations: IIT Jodhpur[1]

Orcid ids: 0000-0002-4781-1500[1]

Contact e-mail: kamansweta@gmail.com

License information: This work has been published open access under Creative Commons Attribution License <http://creativecommons.org/licenses/by/4.0/>, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Conditions, terms of use and publishing policy can be found at <https://www.scienceopen.com/>.

Preprint statement: This article is a preprint and has not been peer-reviewed, under consideration and submitted to ScienceOpen Preprints for open peer review.

DOI: 10.14293/S2199-1006.1.SOR-.PPCV5IA.v1

Preprint first posted online: 24 May 2020

Keywords: Sentiment Analysis, SemEval, LSTM, Deep Learning

Sentiment Analysis on News Articles using Attention Network

Sweta Kaman¹

Department of Science of Intelligene, Indian Institute of Technology Jodhpur, India

kaman.1@iitj.ac.in

Keyword: Sentiment Analysis, Natural Language Processing, Attention, News Classification, SemEval, LSTM, News Sentiment Analyzer.

Abstract: The aim of this paper is to present an approach that can perform a sentiment analysis on any news article with accurate results. For this task I have used Attention Network which is a Deep Learning Framework. There are very few works done in the news sentiment analysis that predicts all the three basic sentiments i.e "positive, negative and neutral" at both sentence-level as well as document-level. The proposed approach is elaborated well in this paper, which gives us a very good accuracy rate for all the three basic sentiments and that can be adopted for any task of sentiment analysis as well.

1. Introduction

News articles and websites are the source of information for all of us which should be accurate and unbiased. These news headlines influences all of us in one or another way. Some of the news headlines gives us positivity, motivation and happiness, while some are negative and create an environment of worries, terror and tensions, and rest of the news does not affects us at all. Also few websites and channels are corrupt and work for specific parties and publish biased news which spread chaos amongst people.

To identify such types of article sources that only spreads negativity can help us to boycott them and sustain a positive environment. Sentiment Analysis [1] helps us to understand the opinions of people and identify such types of intentions. The aim of this project is to detect and classify the news articles as positive, negative and neutral with accurate probabilistic scores by using Attention Network [2] which is a Deep Learning Framework.

2. Dataset Preparation

Knowledge about the suitable type of dataset, as well as manipulation of the unstructured data into a structured format as per our convenience is utmost important if we want our model to predict high accuracy results. The details about the data preparation for both train and test datasets are elaborated in the following subsections.

2.1. Train Dataset(SemEval-2016)

Train Dataset has been prepared by using SemEval-2016 Task 4: Sentiment Analysis on Twitter [3] which consists of training, dev and devtest datasets. I have merged the Training Dataset files and Additional Dataset files into a single Train Dataset file which consists of 53368 sentences.

The train dataset consists of sentiment tags namely positive, negative and neutral corresponding to each of the sentences. I manipulated the dataset by removing the sentiment tags column and by adding three separate columns namely-"Positive, Negative and Neutral" each one which is represented in binary i.e if the sentence tag is positive then the value in the Positive Column will contain binary value 1 and rest of the columns will contain value 0, so on.

2.2. Test Dataset(News Crawler)

To prepare the Test Dataset, I built a web crawler by using a python library i.e BeautifulSoup [4] which helps us to parse through the HTML and XML files. This web crawler will collect all the articles present in the news website and tokenize them into sentences. I also incorporated this web crawler with some preprocessing algorithms that will give us clean sentences. These sentences will be compiled together to form a clean test dataset which will be fed into the sentiment analyser and the output will be predicted probability scores corresponding to each of these sentences.

3. Methods

As soon as the Train Dataset and Test Dataset is prepared by using SemEval-2016 and BeautifulSoup respectively, Glove Embeddings [5] are used as word vector representation which consists of 6B tokens, 400K vocab, uncased, 50d, 100d, 200d and 300d vectors. The dataset is prepared in such a way that it consists of all the sentences of a news article in a sequence which is compiled together along with the whole article. The purpose of incorporating the whole document itself with the compiled sentences of the article is to help us in predicting the document-level probabilistic scores. These scores will help us to detect the intent of the news source. Pre-processing of the dataset is done followed by preparation of embedding matrix and splitting up of dataset into train and val. And then it is fed into the network which uses LSTM merging it with Attention Network. The summary of the model is shown in the following figure.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 150)	0
embedding_1 (Embedding)	(None, 150, 300)	30000000
lstm_1 (LSTM)	(None, 150, 300)	721200
dropout_1 (Dropout)	(None, 150, 300)	0
attention_1 (Attention)	(None, 300)	450
dense_1 (Dense)	(None, 256)	77056
dropout_2 (Dropout)	(None, 256)	0
batch_normalization_1 (Batch Normalization)	(None, 256)	1024
dense_2 (Dense)	(None, 3)	771
Total params: 30,800,501		
Trainable params: 799,989		
Non-trainable params: 30,000,512		

Fig.1 Summary of the Network

4. Experimental Result

After training the model for 25 epochs with batch size of 256 the model obtained an accuracy of 0.9186 and loss of 0.1885 as shown in the following figure. The validation accuracy went up to 0.9257 and the validation loss went down to 0.1948.

```
Epoch 25/25
42694/42694 [=====] - 478s 11ms/step - loss: 0.1885 - accuracy: 0.9186 - val_loss: 0.1948
- val_accuracy: 0.9257
```

Fig.2 Statistics after training the model

4.1. Sentiment Predictions

The final predictions will consists of News sentiment scores for Positive, Negative and Neutral sentiments. The highest amongst the scores will determine the actual predicted sentiment. This contains Both sentence and document level scores where the last line will show the document-level scores and rest of the lines are the sentences from the news article. The predictions are shown in the following figure.

ID	Sentences	positive	negative	neutral
0	IT spending in India is expected to drop by 4....	0.024223	0.001335	0.965741
1	The research firm expects most of this decline...	0.044634	0.023480	0.941933
2	The IDC said that segment would drop by 4.1% d...	0.100878	0.425600	0.524817
3	"Growth of IT services is expected to be almos...	0.015181	0.009076	0.978938
4	Reports say that 90% of IT employees and 70-80...	0.038299	0.308964	0.735429
...	...	...	...	...
305	With the aviation industry in India set for ex...	0.004831	0.032741	0.959809
306	The South Runway will begin with limited landi...	0.136145	0.111380	0.746495
307	The runway will initially have CAT I certifica...	0.044277	0.001340	0.930856
308	Share Via	0.105787	0.149304	0.740689
309	IT spending in India is expected to drop by 4....	0.000238	0.000004	0.999970

Fig.3 Final Predictions for the news article at sentence-level as well as document-level

As depicted in the figure, the probabilistic scores column i.e Positive, Negative and Neutral contains scores and the highest amongst them will help us to identify the sentiment of the sentence and the document. The last line of the result is the document-level prediction of the news article, which shows that the articles possesses a neutral sentiment.

For the code kindly send me an email to the given mail id.

5. Future Work

This sentiment analyser can be used in various tasks such as detecting the sentiments of people during pandemic such as COVID-19, it can also be further manipulated to detect the fake news in various platforms like twitter, facebook, etc. Various Deep Learning models like BERT, XLNET, etc can be used to further increase the accuracy of the model.

References

- [1] B. Pang, L. Lee, *Opinion mining and sentiment analysis, Foundations and trends in information retrieval 2 (1-2) (2008) 1-135.*
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998-6008.*
- [3] P. Nakov, A. Ritter, S. Rosenthal, V. Stoyanov, F. Sebastiani, *SemEval-2016 task 4: Sentiment analysis in Twitter, in: Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval '16, Association for Computational Linguistics, San Diego, California, 2016.*
- [4] G. N. Chandrika, S. Ramasubbareddy, K. Govinda, E. Swetha, *Web scraping for unstructured data over web, in: Embedded Systems and Artificial Intelligence, Springer, 2020, pp. 853-859.*
- [5] J. Pennington, R. Socher, C. D. Manning, *Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532-1543.*
URL <http://www.aclweb.org/anthology/D14-1162>