Nat. Volatiles & Essent. Oils, 2021; 8(6): 4605-4611

NVEO
Natural Volatiles &
Essential Oils

Twitter Data Sets For Disaster Detection And Tracking

Dr. Rupesh Mishra[1] , Dr. Rafael Berlega Llovori[2] , Dr. Kannan Srinathan[3]

[1]Assistant Professor, CBIT, Hyderabad, India

[2] Professor, UJI-I, Spain

[3] Professor, IIIT Hyderabad

Abstract

Nowadays lots of information we can see about disasters either man made or natural, but we don't have any tracking system which can be tracked or identify the disaster in previous before coming to a particular place. Here I have applied the BERT approach of machine learning algorithm which will classify the datasets after building a model with good accuracy of prediction. Also, we have taken datasets from www.iswsm.com website and divide www.iswsm.com disaster based datasets into training and testing sets. After pre processing and tokenized the data and build a model with the help of BERT and CNN deep learning algorithm and trained this model with less loss of data and optimized those models with ADAM to enhance the accuracy and efficiency of the model after deployment in the real life. Compare both models CNN and BERT the accuracy of BERT is high. This system also can be helpful in the area of medical, tourism and weather forecast to give prediction before anything happening.

Keywords: Tensorflow, CNN, Tokenizer, BERT, ADAM, binary_cross entropy, text2bert

Introduction

Disaster detection and tracking are the most important part of the disaster management, because if we know this problem before disasters have come then we can stop too much loss of human, animals, building and different type of resources. Also, we can provide appropriate resources to that place, where disasters are going to happen.

The given datasets for all the information and which type of disasters had happened and going to be happening. For example health issued disaster, flood, Tsunami, Haiti and etc. The main difficulty is to identify about in which place disasters and which type of disasters is going to happen and what datasets should be reprocessed. Because this data should be online and just before of disasters data need to preprocess. One more thing we need in order to keep in mind is that data must be correct and realistic. Noisy data and wrong facts will greatly affect the quality of the results. So, the main challenges are both to identify **high quality data** and mine it properly. Besides, we can regard data from online news agencies, and therefore managing **time issues** is a very Important part about the detection of disasters and its tracking.

In the proposed application, our aim is to retrieve and analyze all instant online data at a given time. Algorithm part should be applied, which have to more efficient for the implementing the tool. Also data structure should be very important role play for developing more efficient and fast tool development. The whole project will be regarded as a Data Science project.

By using Machine Learning algorithm BERT, we will aim at determining which **place** and **incident type (disaster type)** are likely to happen given the captured data. Also here we have used CNN model to find out

the result but we have seen the accuracy of CNN model is poor than BERT model. In this classification, we can know about the content and separate all the content with the help of probabilistic approach. Also we can classify the problems and separately through the two or three dimensional plan for the given datasets.

In this paper, the actual use of online news agencies and social media and its potential towards the disasters detection is a very important role to detect and track the disasters and its location before the disasters happened. In the disaster, the online news agencies (BBC news, TOI, National News, Telegraph, etc) and social media (Facebook, Blogs, twitter, Instagram, Hike and etc) provide online data and that data will process and classify through Machine Learning to identify the proper result which we want to solve the disaster detection and tracking. Also different types of problem will be created if we have taken data from the social media data, because may be any persons if gave wrong post regarding the disasters then our system will execute that data also. So be cautious about that type of data. This will be the challenge how to ignore and remove that data from the social media during the design of our system.

Dataset description

Disaster relief is the .csv file, which is taken from the website www.iswsm.com. In which, which type of disasters had happened and what type of resources and help had been provided, all details have given in the disaster-relief-dfe-854578.csv file. This data come from twitter datasets in which different types of labels with tagline has been mentioned.

Datasets are related to the paper [2], which is of taken in a file of 9300 # observations. Many more papers has been used this datasets. Most of the file we are not getting to access because of security code and password. Only one file out of seven links has given permission to access.

Related Works

Methodology

We have proposed steps for preprocessing content of datasets of disasters. Sequence has been maintained for find out the detection and tracking of the disasters. The sequences are follows

i.) Collection of all tweeter datasets in disasters relief .csv files from one of the website.

ii.) Pre-process all the content inside the csv file and remove all noisy data from it.

iii.) Split training and testing datasets of the given unorganized data from Twitter data sets.

iv.) Apply **BERT Model** representation model to train the model and test on it.

vii.) Apply LSTM upon BERT to find out regression value.

viii.) Find out the accuracy of this model after applying formulae.

ix.) Based upon the accuracy of the designed model, disasters can be detected and track in previous before happening.

x.) Compare the accuracy of LSTM with CNN Deep learning Algorithm.

Here, we have introduced three flow diagram related to tourist places for achieving the goal. We have designed the combined CNN and LSTM model of three opinion based tokens. Here we can see when we are comparing accuracy of both models then accuracy of LSTM model is good. Also CNN model is very complex to implement but LSTM is too easy to implement. Using a trained sets and test set upon the tourism datasets applies tensor flow and a caress Machine learning algorithm to build a model. We have tokenized all the retrieved data sets with the help of the to kenizer techniques. Apply LSTM techniques to train the model. Applying optimizer techniques to know how much loss we have when make this model.

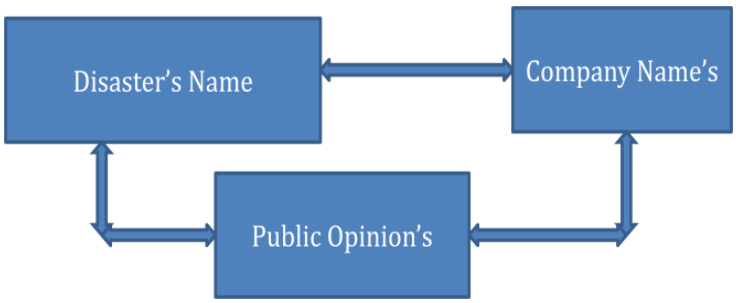

Figure 1: Three semantic perspective for characterizing performance context.

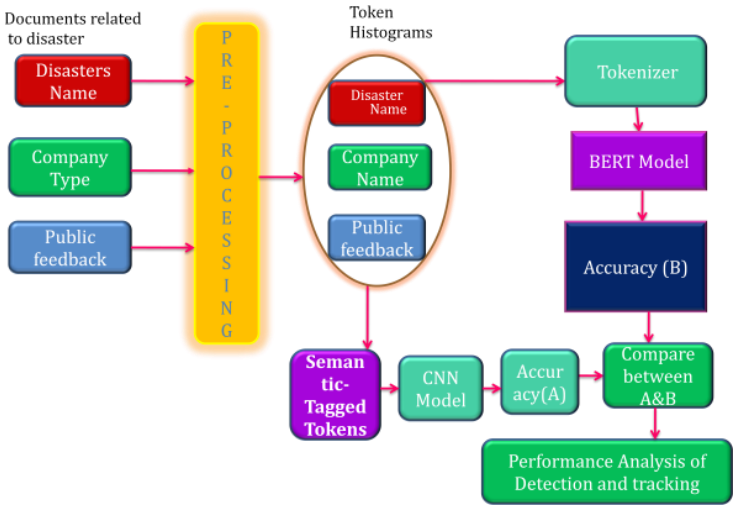

Figure 2: Working model of CNN, LSTM and BERT

Evaluation and Results

Here we have taken datasets disaster-relief- dfe-854578.csv of natural disasters from [19]. After preprocessing all datasets we have converted it into trained and test sets with text and label of those text.

	text
3651	Las Vegas in top 5 cities for red-light runnin...
6118	Do you feel like you are sinking in unhappines...
7147	The Architect Behind Kanye West□Û³s Volcano ht...
4669	@ZachLowe_NBA there are a few reasons for that...
5204	I can't wait to be beyond obliterated this wee...

Figure1: format of disaster datasets after label

Based upon the datasets we have formed train sets and test based after give label for the particular text information related to disasters.

Evaluate the model

```
In [51]: result, model_outputs, wrong_predictions = model.eval_model(vali

In [52]: result

Out[52]: {'eval_loss': 0.4463166512640102,
          'fn': 200,
          'fp': 165,
          'mcc': 0.6727919417893213,
          'tn': 1138,
          'tp': 781}
```

Figure 2: Evaluation of the model

After evaluation of the model we can see the loss is of 44% and we need to reduce the loss, then here we have used optimizer techniques called 'ADAM' and loss function is called 'binary_crossentropy' to enhance the accuracy of the model after BERT techniques.

Here we have found the accuracy of the CNN model is 81% and LSTM model is 93% as compared this model about how much accuracy has been given when we have trained our model.

```
predictions = []
for x in model_outputs:
    predictions.append(np.argmax(x))

print('f1 score:', f1_score(valid_df['labels'], predictions))
```

```
f1 score: 0.8105864037363778
```

Figure 3: Accuracy of the CNN model

In the below images we can see that after applying optimizer techniques 'ADAM' to find out the loss of this model.

```
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

```
model.fit(
    [  np.array(input_ids_all),  np.array(input_masks_all), np.array(input_segments_all) ],
    np.array(classes),
    validation_split = 0.3,
    epochs = 4,
    batch_size = 50
)
```

```
Epoch 1/4
2234/2234 [==============================] - 744s 333ms/step - loss: 0.2151
y: 0.9247
Epoch 2/4
2234/2234 [==============================] - 743s 333ms/step - loss: 0.1885
y: 0.9316
Epoch 3/4
2234/2234 [==============================] - 742s 332ms/step - loss: 0.1820
y: 0.9298
Epoch 4/4
2234/2234 [==============================] - 741s 332ms/step - loss: 0.1796
y: 0.9300
```

Figure 4: Accuracy of the LSTM model

Conclusion

In this paper, we have taken datasets from different online websites and twitter and apply different deep learning and machine learning algorithm for finding out better accuracy of the model. After application of CNN model and BERT model, we have seen that CNN model is good but not easy to implement, but in the case of BERT it's easy to use and given better accuracy upon our datasets. Whenever we have compared BERT with CNN then we can see the accuracy of BERT model is good as compare with CNN. In the future, these types of model can be used for medical purposes, tourism industry and automobile industry.

REFERENCES

David E. Alexander "Social Media in Disaster Risk Reduction and Crisis Management", Springer Science+Business Media Dordrecht 2013

Y He, C Lin, W Gao, and KF Wong."Tracking Sentiment and Topic Dynamics from Social Media".

G. Moya, L., Berlanga, R., Nebot, V.,Aramburu, M.J.,Llido, D.,Sanz, I. "An Open Data Infrastructure for Enabling Social Business Intelligence", International Journal of Data Warehousing and Mining, 11(4)**,** 1-28, October-December 2015.

E. Schnebele , G. Cervone , S. Kumar and N. Waters, "Real Time Estimation of the Calgary Floods Using Limited Remote Sensing Data" , Water 2014, 6(2), 381-398.

S. Kumar, X. Hu, H. Liu, "A behavior analytics approach to identifying tweets from crisis regions", ISBN: 978-1-4503- 2954-5, ACM-2014.

B. Mandel, A. Culotta, J. Boulahanis, Danielle Stark, Bonnie Lewis, Jeremy Rodrigue, "A Demographic Analysis Of Online Sentiment during Hurricane Irene",2012.

A. Nagy, J. Stamberger, "Crowd Sentiment Detection during Disasters and Crises", Proceedings of the 9th International ISCRAM Conference-Vancovuver, Canada, April 2012.

R. Mishra, K.Saini (2014), "Automatic Detection of Interlinked Events for Better Disaster management", IEEE, IACC-2014, ITM Gurgaon.

T. Mikolov, I. Sutskever, K. Chen, G. Carrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality" NIPS, pages 3111- 3119, 2013

T. Mikolov "Distributed Representation of Sentences and Documents", International conference on machine learning, Beijing, China, 2014, JMLR: W&CP volume 32.

T. Mikolov, "Efficient Estimation of Word Representation in Vector Space".

CoRR, abs/1309.4168, 2013b

Y. Kim, "Convolutional neural network for sentence classification", arXiv:1408.5882v2, 2014.

J. Weston, A. Borders, O. Yakhnenko and N. Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In proceeding of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Morin, Frederic and Bengio, Yoshua. Hierarchical probabilistic neural network language model. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, pp. 246–252, 2005.

Pang, Bo and Lee, Lillian. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In Proceedings of Association for Computational Linguistics, pp. 115–124, 2005.

Perronnin, Florent and Dance, Christopher. Fisher kernels on visual vocabularies for image categorization. In IEEE Conference on Computer Vision and Pattern Recognition, 2007.

Perronnin, Florent, Liu, Yan, Sanchez, Jorge, and Poirier, Herve. Large-scale image retrieval with compressed fisher vectors. In IEEE Conference on Computer Vision and Pattern Recognition, 2010.

Rumelhart, David E, Hinton, Geoffrey E, and Williams, Ronald J. Learning representations by back-propagating errors. Nature, 323(6088):533–536, 1986.

Socher, Richard, Huang, Eric H., Pennington, Jeffrey, Manning, Chris D., and Ng, Andrew Y. Dynamic pooling and unfolding recursive auto encoders for paraphrase detection. In Advances in Neural Information Processing Systems, 2011a.

Socher, Richard, Lin, Cliff C, Ng, Andrew, and Manning, Chris. Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML11), pp. 129–136, 2011b.

Socher, Richard, Pennington, Jeffrey, Huang, Eric H, Ng, Andrew Y, and Manning, Christopher D. Semisupervised recursive autoencoders for predicting sentiment distributions. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011c.

Socher, Richard, Chen, Danqi, Manning, Christopher D., and Ng, Andrew Y. Reasoning with neural tensor networks for knowledge base completion. In Advances in Neural Information Processing Systems, 2013a.

Socher, Richard, Perelygin, Alex, Wu, Jean Y., Chuang, Jason, Manning, Christopher D., Ng, Andrew Y., and Potts, Christopher. Recursive deep models for semantic compositionality over a sentiment treebank. In Conference on Empirical Methods in Natural Language Processing, 2013b.

Srivastava, Nitish, Salakhutdinov, Ruslan, and Hinton, Geoffrey. Modeling documents with deep boltzmann machines. In Uncertainty in Artificial Intelligence, 2013.

Turian, Joseph, Ratinov, Lev, and Bengio, Yoshua. Word representations: a simple and general method for semisupervised learning. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 384–394. Association for Computational Linguistics, 2010.

Turney, Peter D. and Pantel, Patrick. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research, 2010.

Wang, Sida and Manning, Chris D. Baselines and bigrams: Simple, good sentiment and text classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, 2012.

Yessenalina, Ainur and Cardie, Claire. Compositional matrix-space models for sentiment analysis. In Conference on Empirical Methods in Natural Language Processing, 2011.

Zanzotto, Fabio, Korkontzelos, Ioannis, Fallucchi, Francesca, and Manandhar, Suresh. Estimating linear models for compositional distributional semantics. In COLING, 2010.

Zhila, A., Yih, W.T., Meek, C., Zweig, G., and Mikolov, T. Combining heterogeneous models for measuring relational similarity. In NAACL HLT, 2013.

Zou, Will, Socher, Richard, Cer, Daniel, and Manning, Christopher. Bilingual word embeddings for phrasebased machine translation. In Conference on Empirical Methods in Natural Language Processing, 2013.

G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE transactions on knowledge and data engineering, 17(6):734–749, 2005.