Real / Fake Job Prediction

Dina Bishr & Mariam Daabis

Deep Learning, Computer Engineering,The American University in Cairo

Introduction

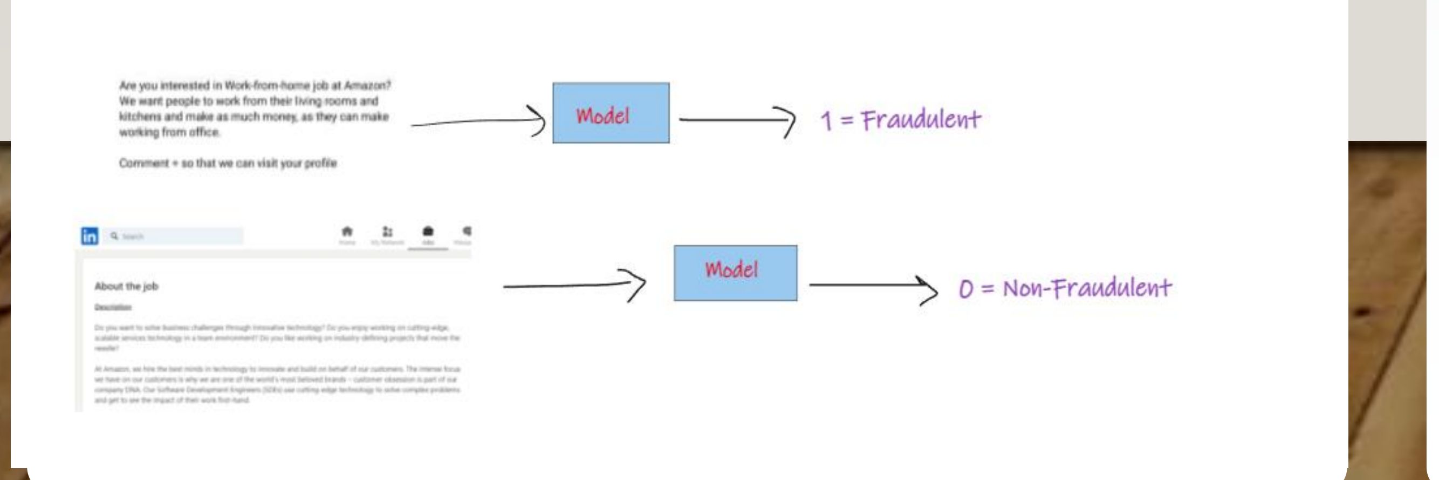
- There has been a significant increase in the number of online jobs offered on various employment portals.
- It has been reported that not all job postings are legitimate
- This pose problems with the job posting website and their credibility and the probability that the user would return to the website.
- It can also be a security threat since the scammers can be using the applicants' information to steal their identities.
- Using advanced deep learning techniques, we are trying to predict whether these job postings are real or fake to be filtered early on.

Dataset

- Our problem contains only one available imbalanced dataset provided by The University of the Aegean and available on Kaggle
- It was used in many models and research papers
- The dataset consists of 18 columns including:

#	Variable	Datatype	Description
1	job_id	int	Identification number given to each job posting
2	title	text	A name that describes the position or job
3	location	text	Information about where the job is located
4	department	text	Information about the department this job is offered by
5	salary_range	text	Expected salary range
6	company_profile	text	Information about the company
7	description	text	A brief description about the position offered
8	requirements	text	Pre-requisites to qualify for the job
9	benefits	text	Benefits provided by the job
10	telecommuting	boolean	Is work from home or remote work allowed
11	has_company_logo	boolean	Does the job posting have a company logo
12	has_questions	boolean	Does the job posting have any questions
13	employment_type	text	5 categories – Full-time, part-time, contract, temporary and other
14	required_experience	text	Can be – Internship, Entry Level, Associate, Mid-senior level, Director, Executive or Not Applicable
15	required_education	text	Can be – Bachelor's degree, high school degree, unspecified, associate degree, master's degree, certification, some college coursework, professional, some high school coursework, vocational
16	Industry	text	The industry the job posting is relevant to
17	Function	text	The umbrella term to determining a job's functionality
18	Fraudulent	boolean	The target variable 0: Real, 1: Fake

Input/output



Are you interested in Work-from-home job at Amazon? We want people to work from their living rooms and kitchens and make as much money, as they can make working from office.

Comment + so that we can visit your profile

About the job

Description

Models

Since this dataset contains job ads which essentially are text, then some Natural Language Processing was performed, and thus the available models are concerned with these types of data.

The models are :

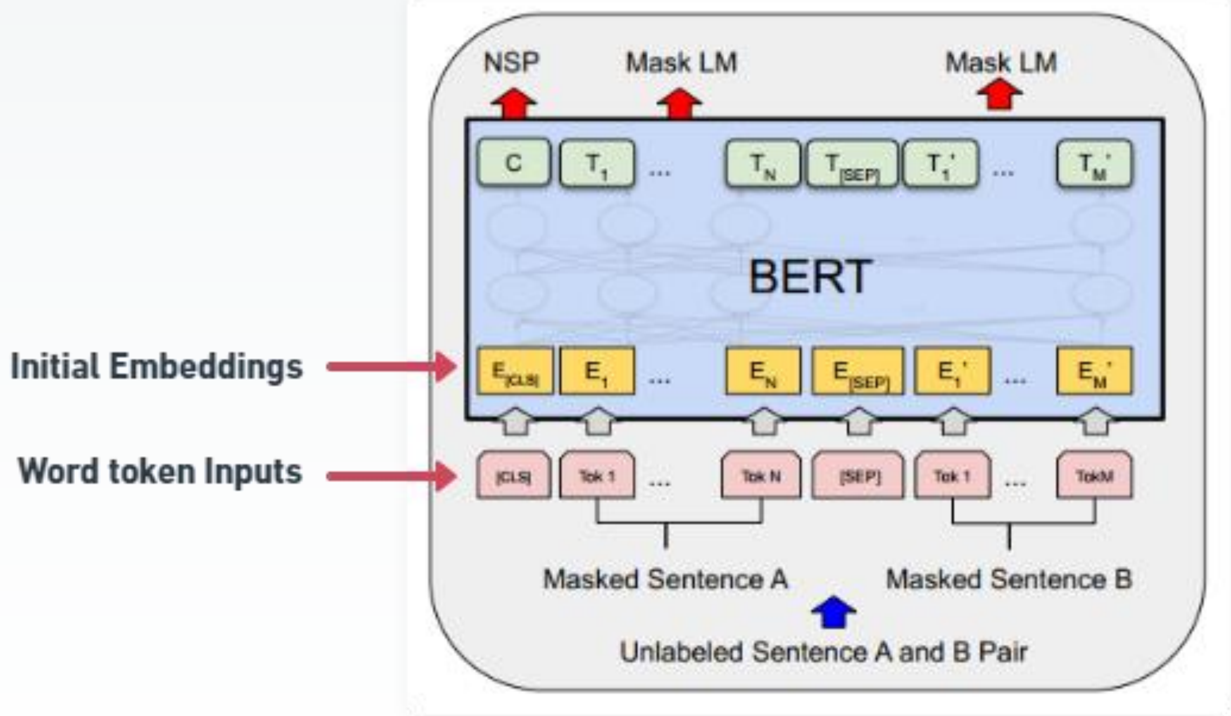
Long Short-Term Memory (LSTM) (Chosen Model)

- Is an artificial recurrent neural network architecture (RNN)
- It is used in deep learning and is capable of learning long-term dependencies
- they are designed to specifically avoid the long-term dependency problem
- They have a chain-like structure where the repeating module has four neural network layers that interact instead of just one
- it does the following: it decides which information it's going to remove from the cell state.
- It does this through a sigmoid layer called the "forget gate layer" that outputs a number between 0 and 1 for each number in cell state where 1 represents "keep" while 0 represents "remove".

BERT

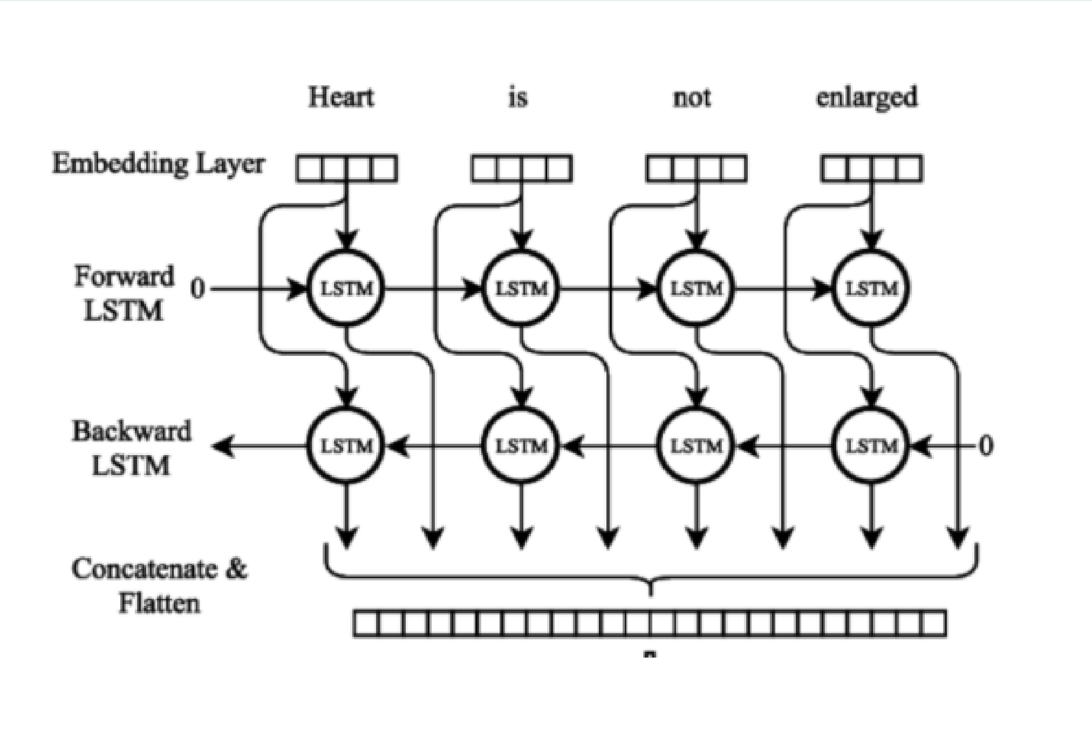
- BERT stands for Bidirectional Encoder Representations
- based on the Transformers library
- a deep learning model is one that contains one or more layers in which every input is connected to every output, hence the reason this model took some time to train.
- is capable of processing text both left-to-right and right-to-left at the same time.
- provides pre-trained models on large amounts of data to facilitate tasks such as semantic labeling and, more importantly, in our case, sentence classification.
- It is able to determine whether two pieces of text possess a connection or are simply unrelated.

iq.opengenus.org

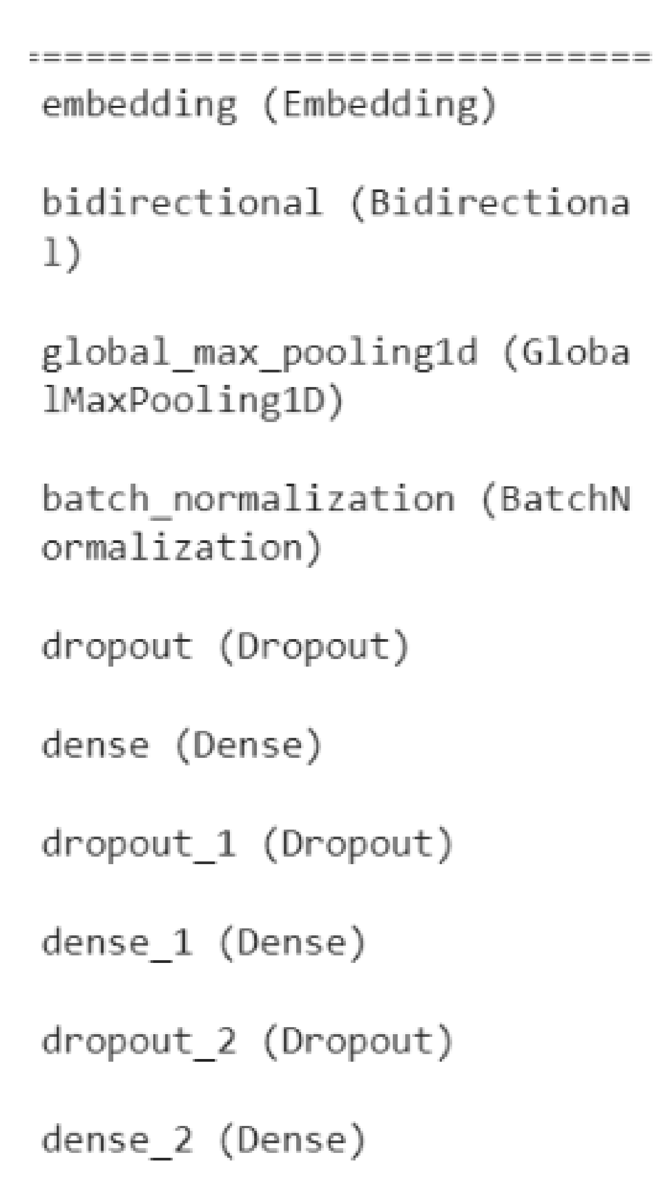


Discussion

- We chose the Bidirectional LSTM model over the BERT
- we are aiming to improve its accuracy and make it closer to the accuracy of the BERT model.
- As shown below, the bi-directional LSTM model consists of two LSTM's: one takes the input in a forward direction, and the other takes the input in a backward direction. BiLSTMs effectively increase the quantity of data available to the network, giving the algorithm better context.



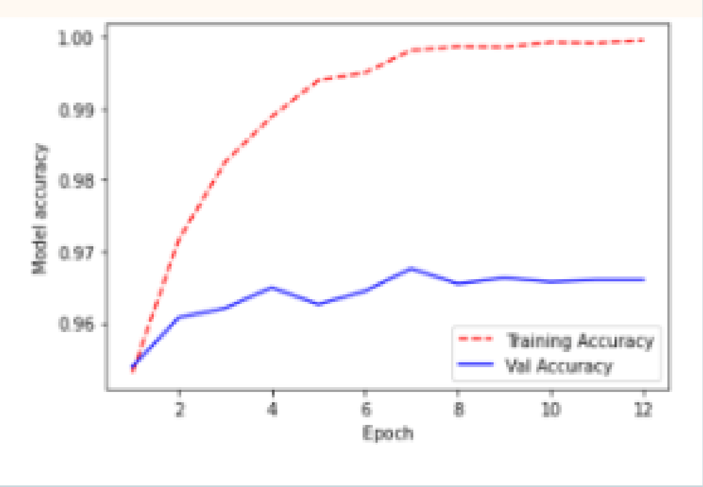
- We adopted the idea of tokenization from Bert model to our model and worked on the model layers to become the following:



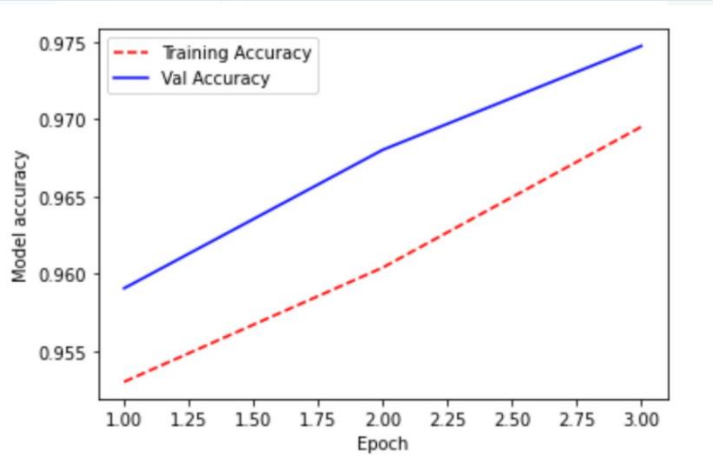
```
===============================
 embedding (Embedding)

 bidirectional (Bidirectiona
 l)

 global_max_pooling1d (Globa
 lMaxPooling1D)

 batch_normalization (BatchN
 ormalization)

 dropout (Dropout)

 dense (Dense)

 dropout_1 (Dropout)

 dense_1 (Dense)

 dropout_2 (Dropout)

 dense_2 (Dense)
```

Results

Initial Result - Imbalanced Dataset



Final Result – Oversampled Dataset



Conclusion

- Not all techniques that solve overfitting work
  - We have tried oversampling the data and it still overfit
- Since the model takes a lot of time to train it was very difficult to do extensive hyper-parameter tuning and variations in the model architecture.
- GloVe and the cleaning of the text and NLP in general highly affected the data thus the results were extremely different in a good way, so we recommend it.

References

- https://arxiv.org/pdf/1810.04805v2.pdf
- https://www.kaggle.com/niduttnb/bert-vs-birdirectional-lstm/notebook#3)-BI-DIRECTIONAL-LSTM
- https://paperswithcode.com/method/bilstm
- https://www.kaggle.com/gauravsahani/real-or-fake-job-postings-with-bi-directional-lstm/notebook
- https://github.com/Anshupriya2694/Fake-Job-Posting-Prediction
- https://paperswithcode.com/method/berthttps://link.springer.com/content/pdf/10.1007/s11063-021-10727-z.pdf
- https://www.kaggle.com/shivamb/real-or-fake-fake-jobposting-prediction