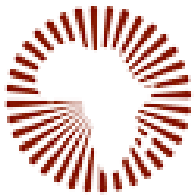
Text Classification Using Fully-connected (FC) NNs

Fadila, Fatoumata , Fonteh, Sakayo

African Master of
Machine Intelligence

April 29, 2022



AIMS | African Institute for Mathematical Sciences
NEXT EINSTEIN INITIATIVE

Outline

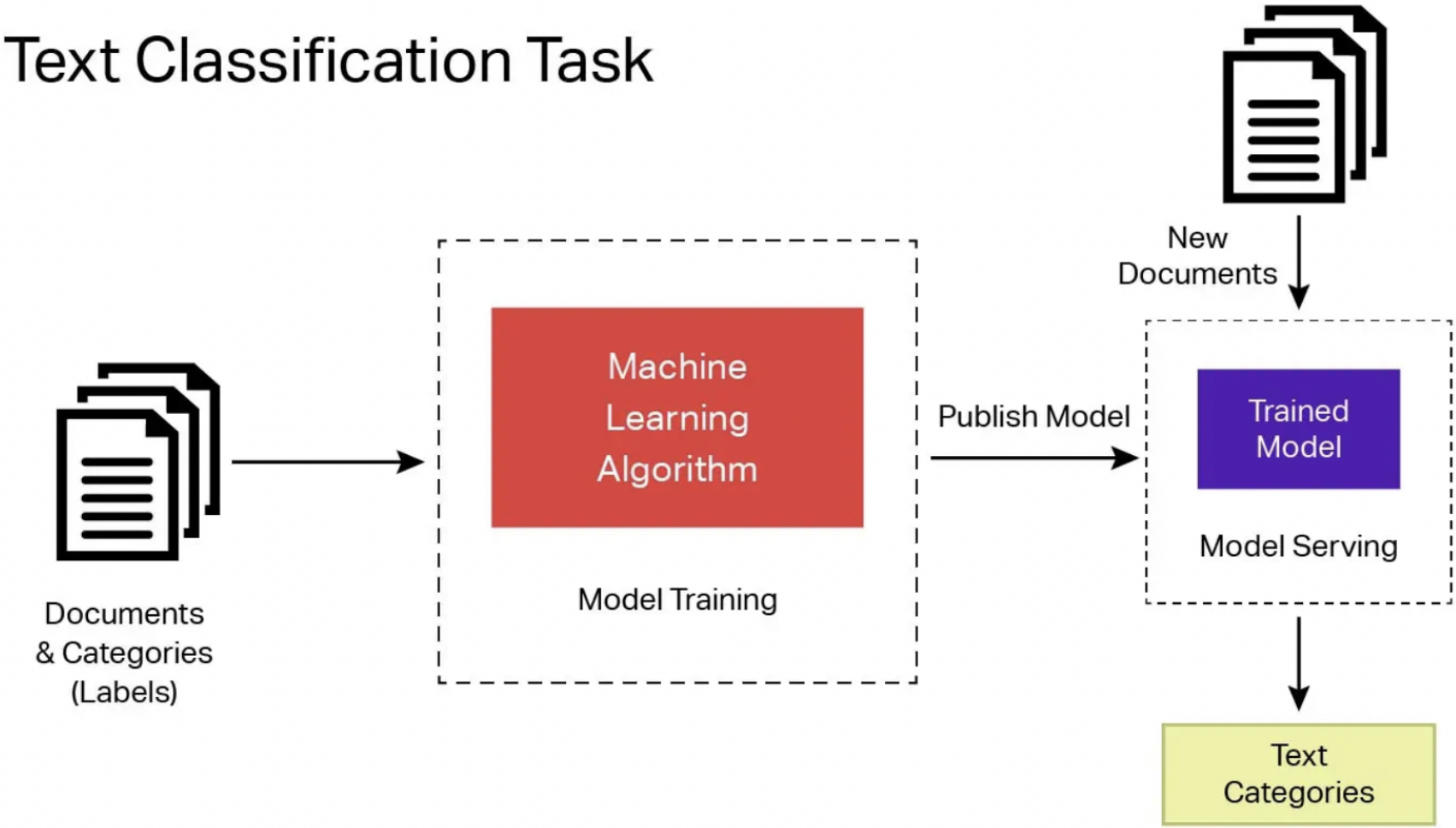
Introduction

Text can be an extremely rich source of information, but extracting insights from it can be hard and time-consuming, due to its unstructured nature.

However, natural language processing and machine learning, which both fall under the vast umbrella of artificial intelligence, has made sorting text data is getting easier.

AIMS African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

Introduction



Methodology

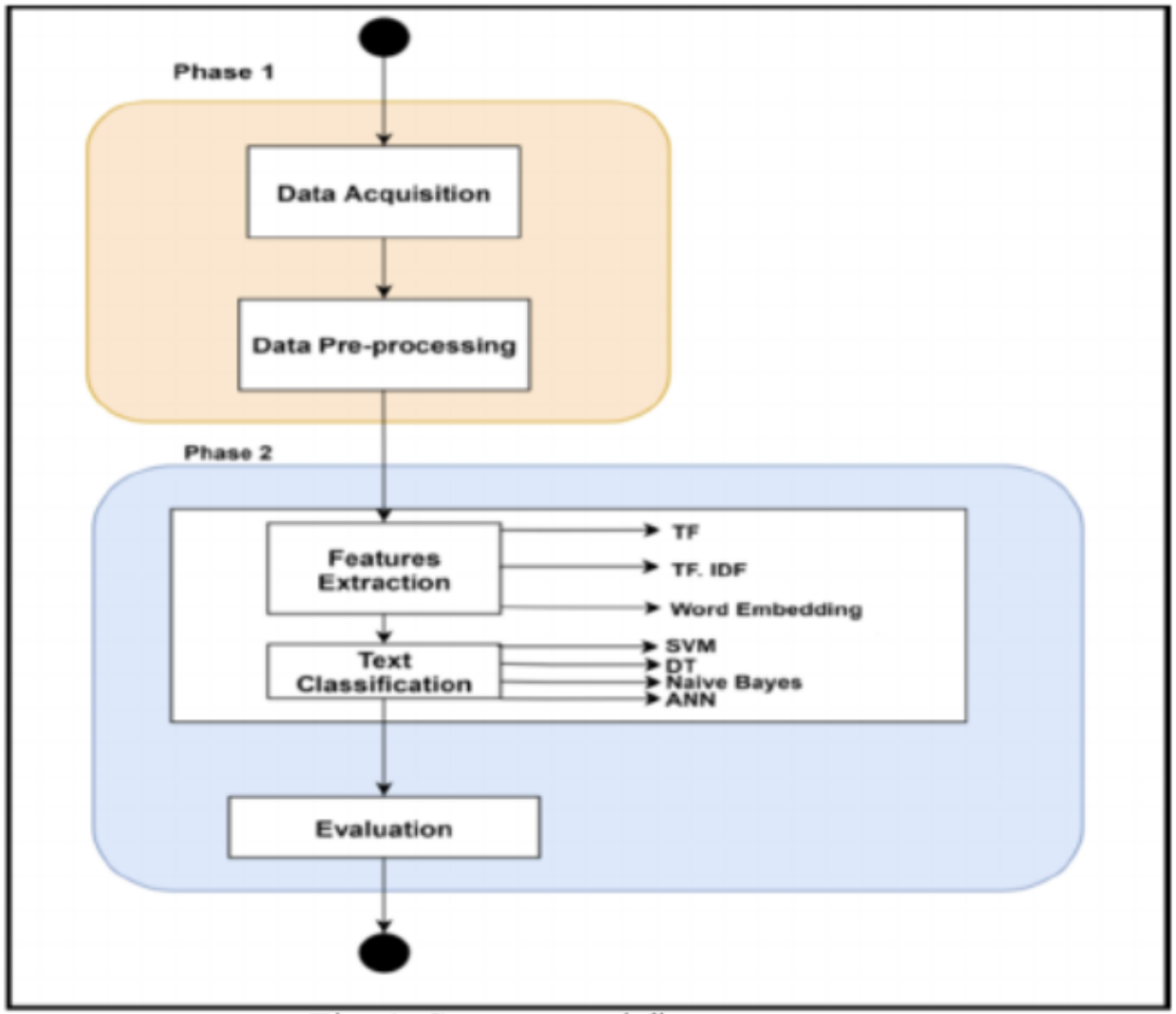
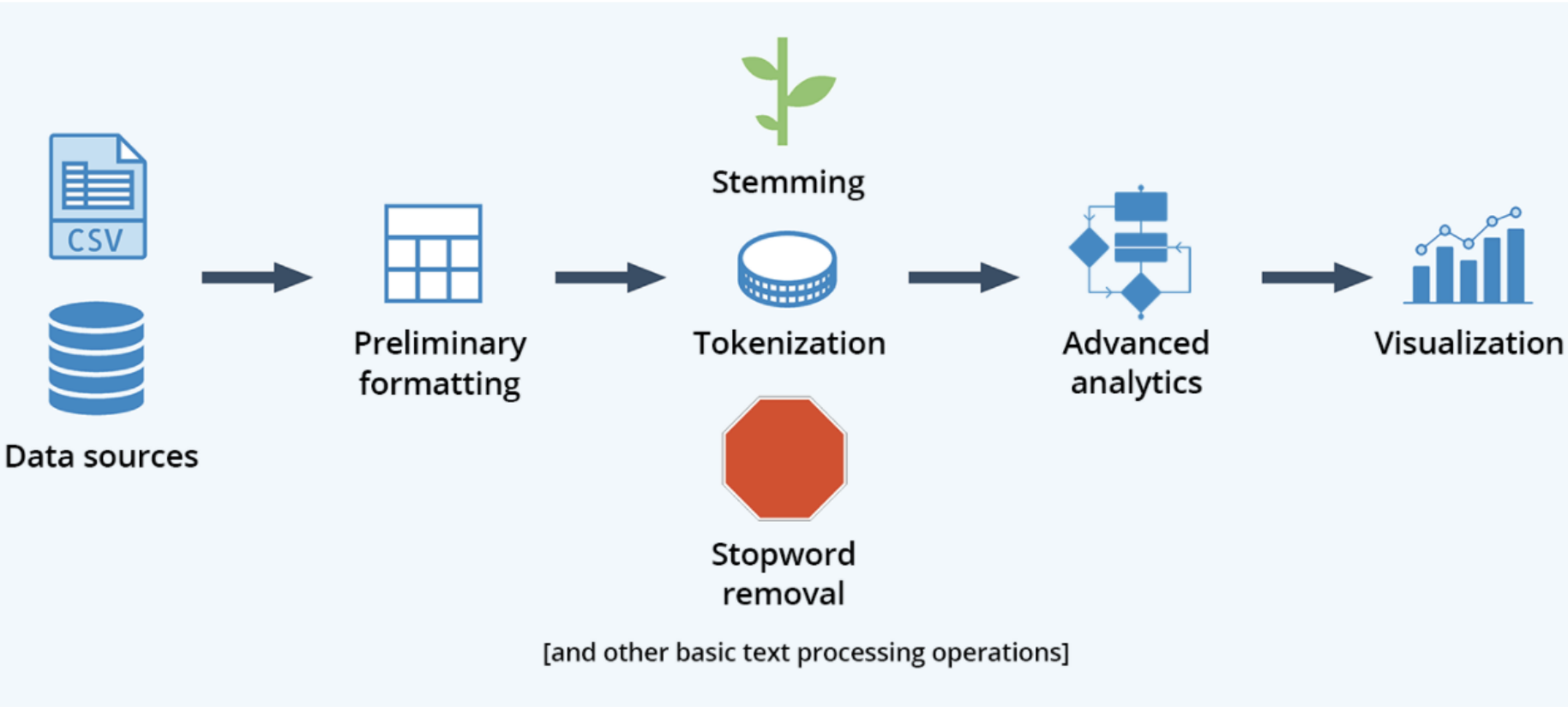


Fig. 1. System workflow.

Data Preprocessing



Sentence Representation

Definition

Sentence representation follow similar representation as word. However, sentences are compound than words and hence uses the following preprocessing techniques:

- Padding
- Truncating

One-hot encoding

Let the vocabulary $V = \{w_1, w_2, w_3, ..., w_{|V|}\}$ every term w in vocabulary.

$$w_i = \begin{cases} 1 & \text{si } w = w_i \\ 0 & \text{otherwise} \end{cases}$$

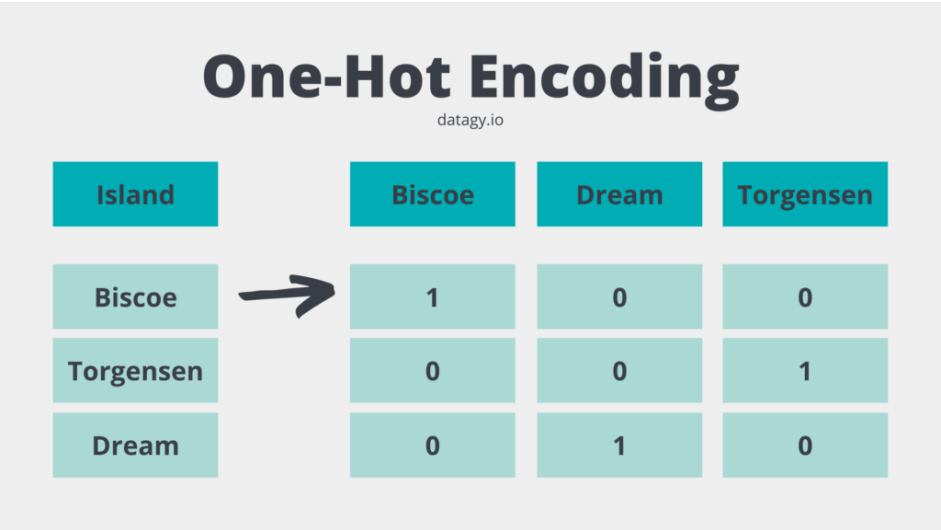


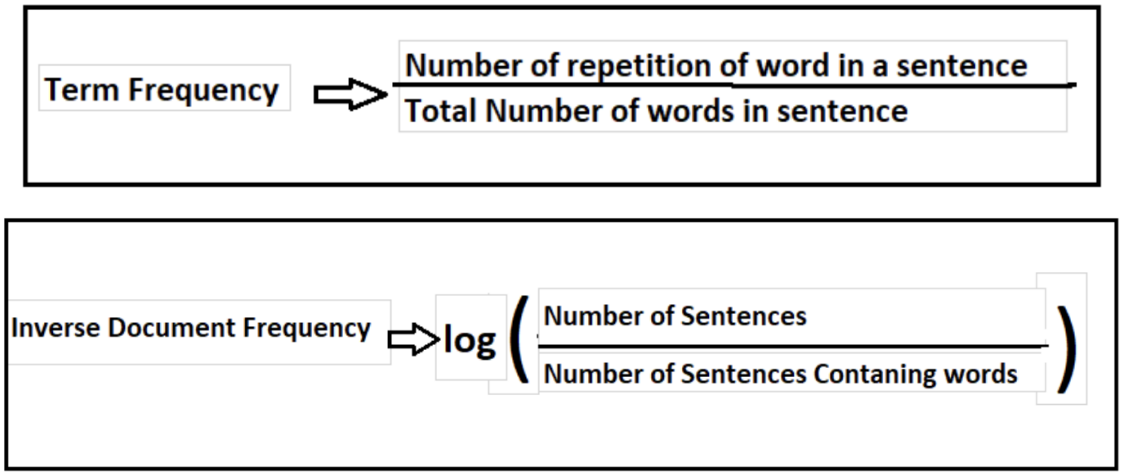
Figure: One Hot encoding

- The size of the vector is equal to the count of unique words in the vocabulary and not convey information about the context.
- It is not capture the relationships between different words.

Term Frequency - Inverse Document Frequency(TF-IDF)

For a term i in document j:

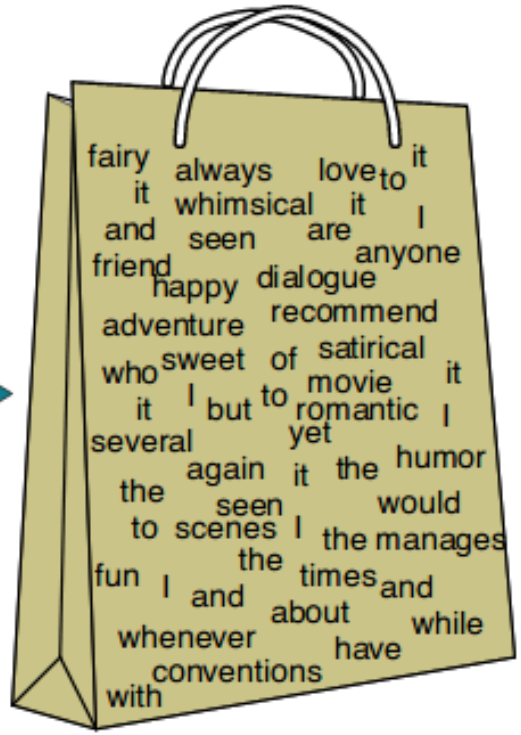
$$TFIDF = TF_{i,j} \times log \frac{N}{DF_i}$$



- It assumes that the counts of different words provide independent evidence of similarity.
- It makes no use of semantic similarities between words

Bag of word

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



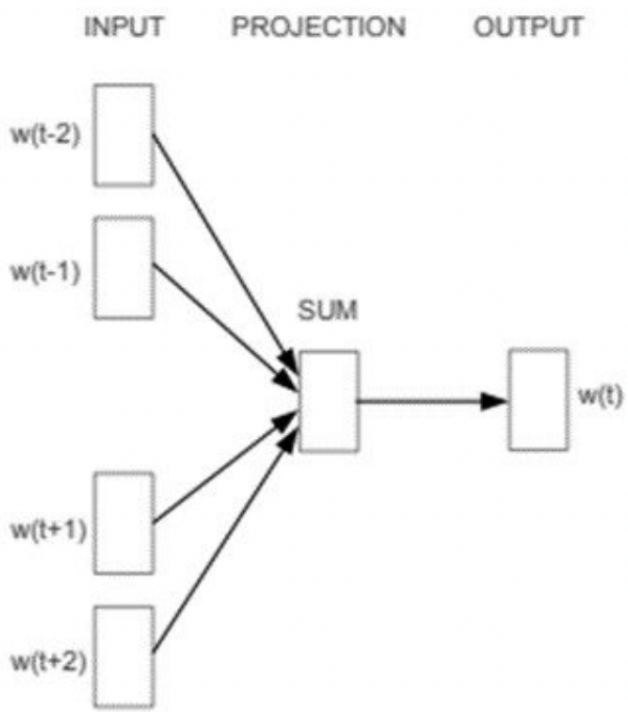
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

- This method doesn't preserve the word order.
- It does not allow to draw of useful inferences for downstream NLP tasks.

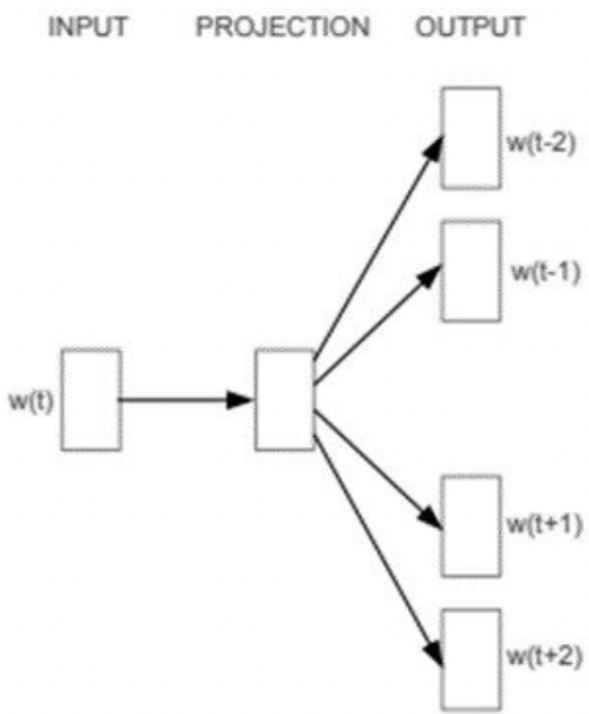
AIMS African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

Word2Vec embeddings

It finds similarities among words by using the cosine similarity metric.It offers two neural network-based variants: Continuous Bag of Words (CBOW) and Skip-gram.



CBOW



Skip-gram

IMS African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

GloVe — Global Vectors for Word Representation

GloVe extends the work of Word2Vec to capture global contextual information in a text corpus by calculating a global word-word co-occurrence matrix.

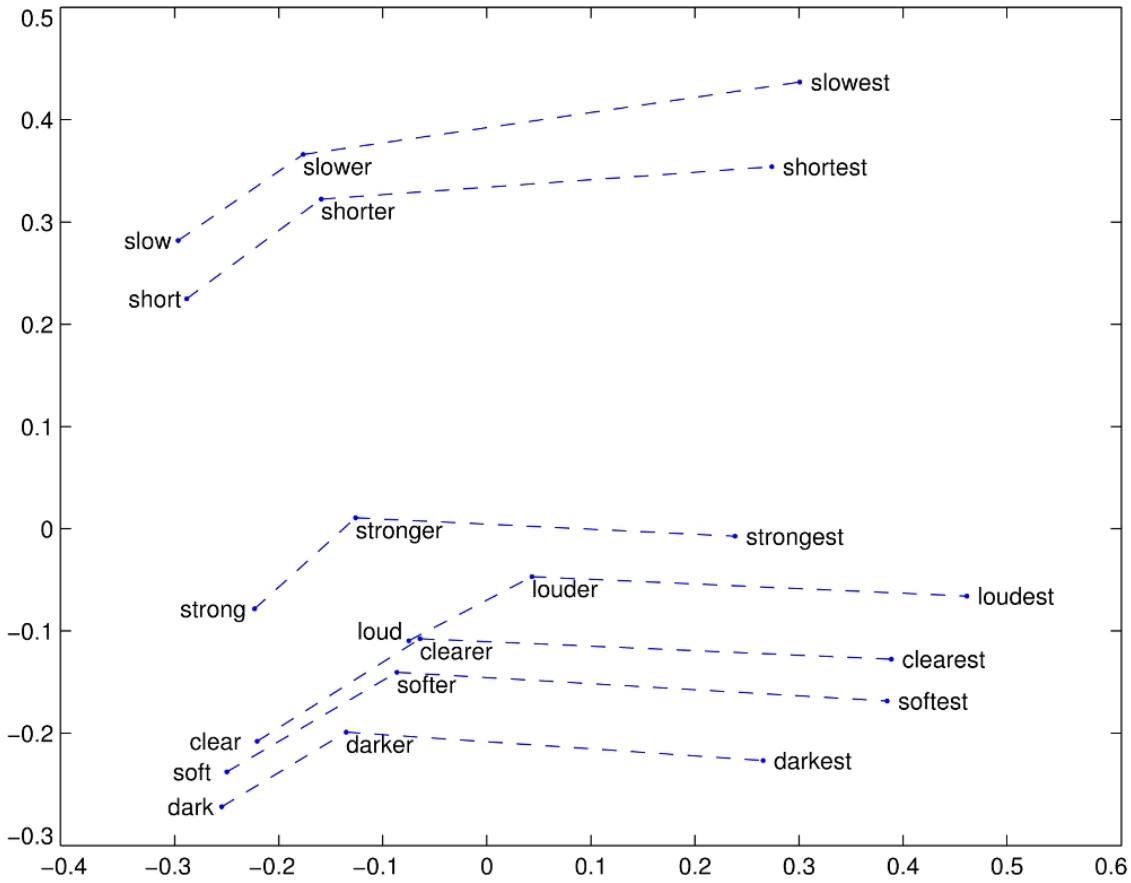


Figure: Glove vectors representations

Fully-connected Neural Networks(FC-NNs)

FCNNs consists of a series of fully connected layers in such that all the nodes, or neurons, in one layer are connected to every neurons in the next layer.

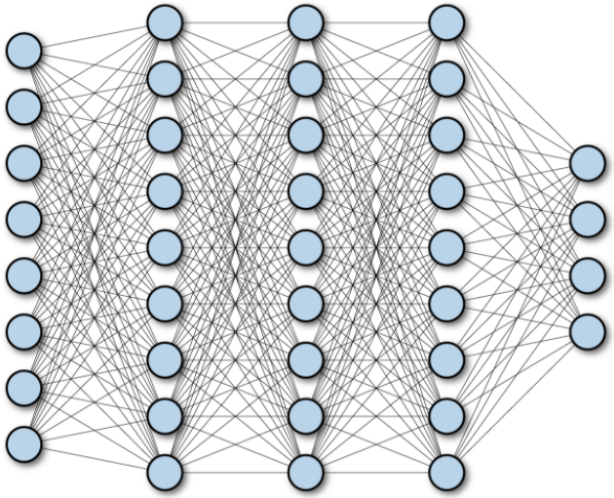


Figure: Multilayer Deep Fully Connected Network

Architecture

In this project, our FCNN have input layer, and two hidden layers(first:100units,second:50units) and an output layer.

Model overview

Definition

Let $x \in \mathbb{R}^m$ represent the input to a fully connected layer.Let $y_i \in \mathbb{R}$ be the i-th output from the fully connected layer. Then $y_i \in \mathbb{R}$ is computed as follows:

$$y_i = \sigma\left(w_1 x_1 + w_2 x_2 + \ldots + w_m x_m\right)$$

Here σ is a nonlinear function and the w_i are weights in the network. The full output y is then

$$y = \begin{pmatrix} \sigma\left(w_{1,1}x_1 + w_2 x_2 + \right. & \ldots & + w_{1,m}x_m \\ \vdots & \ddots & \vdots \\ \sigma\left(w_{n,1}x_1 + w_2 x_2 + \right. & \ldots & + w_{m,m}x_m \end{pmatrix}$$

FCNN forward Pass and Learning

Layer 1 affine:

$$x_1 = W_1 X + b_1$$

Layer 1 activation:

$$h_1 = Relu(x_1)$$

Layer 2 affine:

$$x_2 = W_2 h_1 + b_2$$

output:

$$p = softmax(x_2)$$

Loss:

$$L = -\Big(ylog(p) + (1-y)log(1-p)\Big)$$

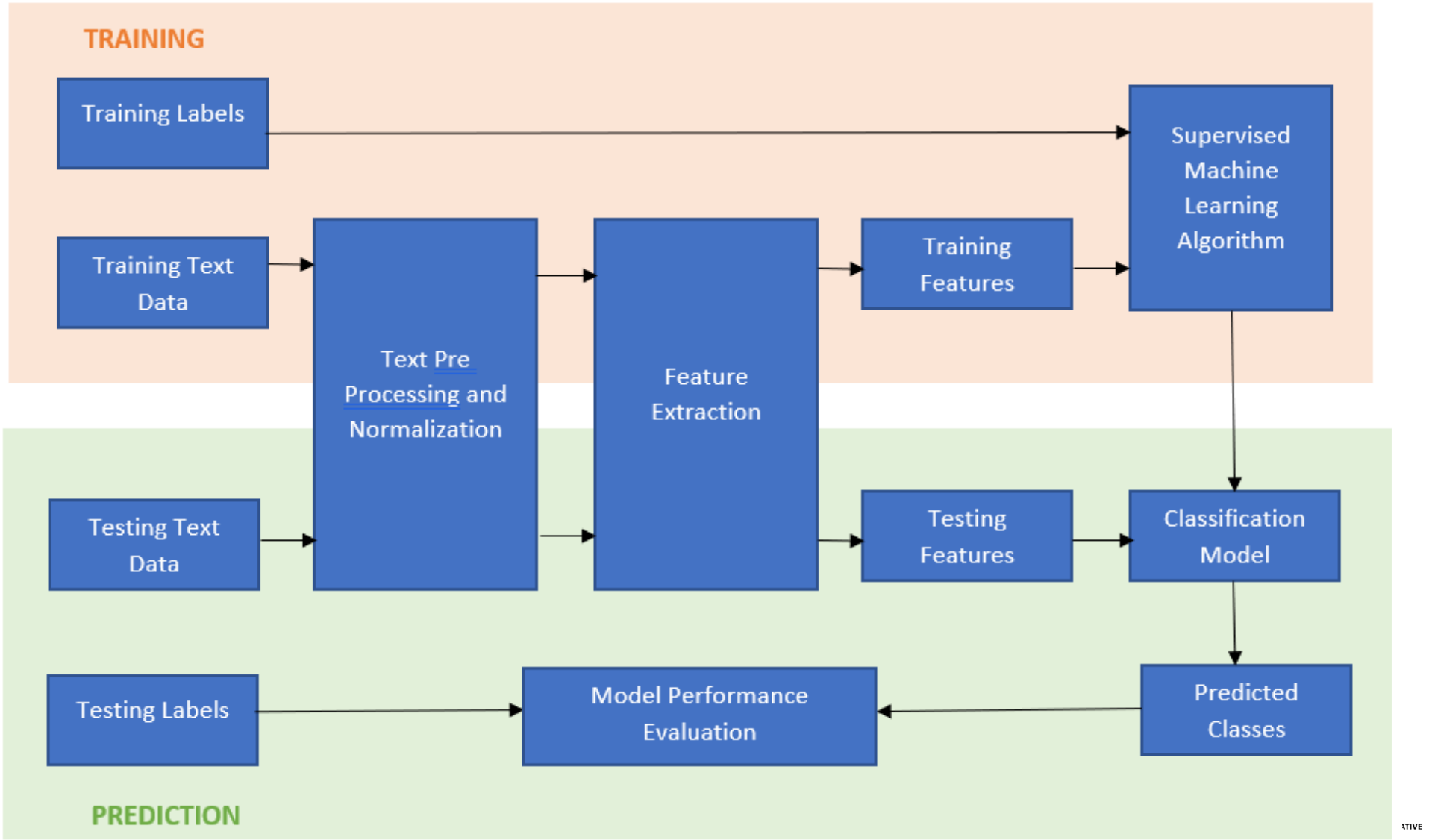
Gradient:

$$\frac{\partial}{\partial W_1} L(W_1, b_1, W_2, b_2) = \frac{\partial L}{\partial p}\frac{\partial p}{\partial x_2}\frac{\partial x_2}{\partial h_1}\frac{\partial h_1}{\partial x_1}\frac{\partial x_1}{\partial W_1}$$

Parameter update: $W_1 = W_1 - \alpha \frac{\partial L}{\partial W_1}$, where α is the learning rate

AIMS African Institute for Mathematical Sciences NEXT EINSTEIN INITIATIVE

Flowchart for Training & Testing



Model with Bag of words

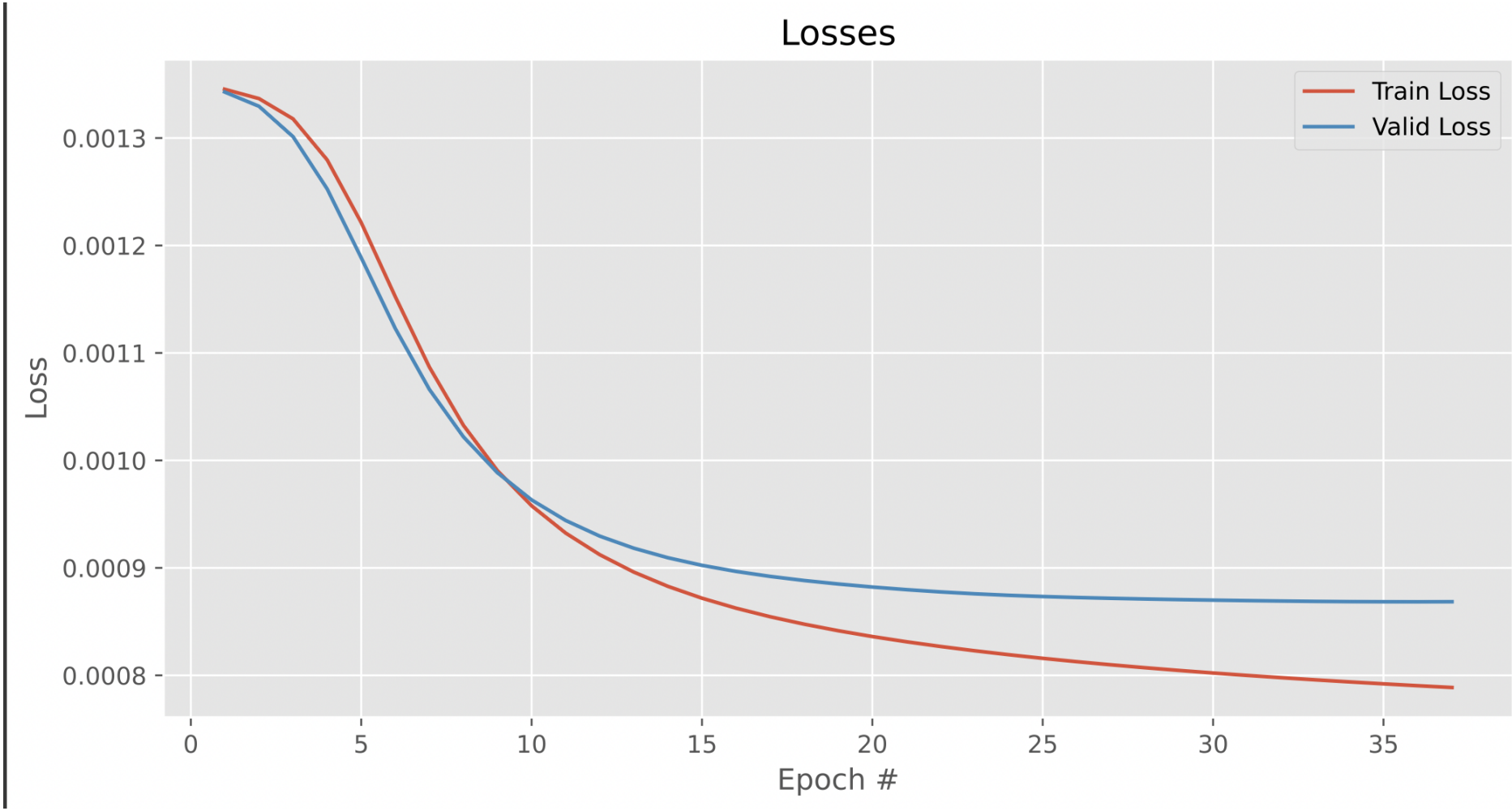


Figure: Training and Validation Losses plot

Model with TF-IDF

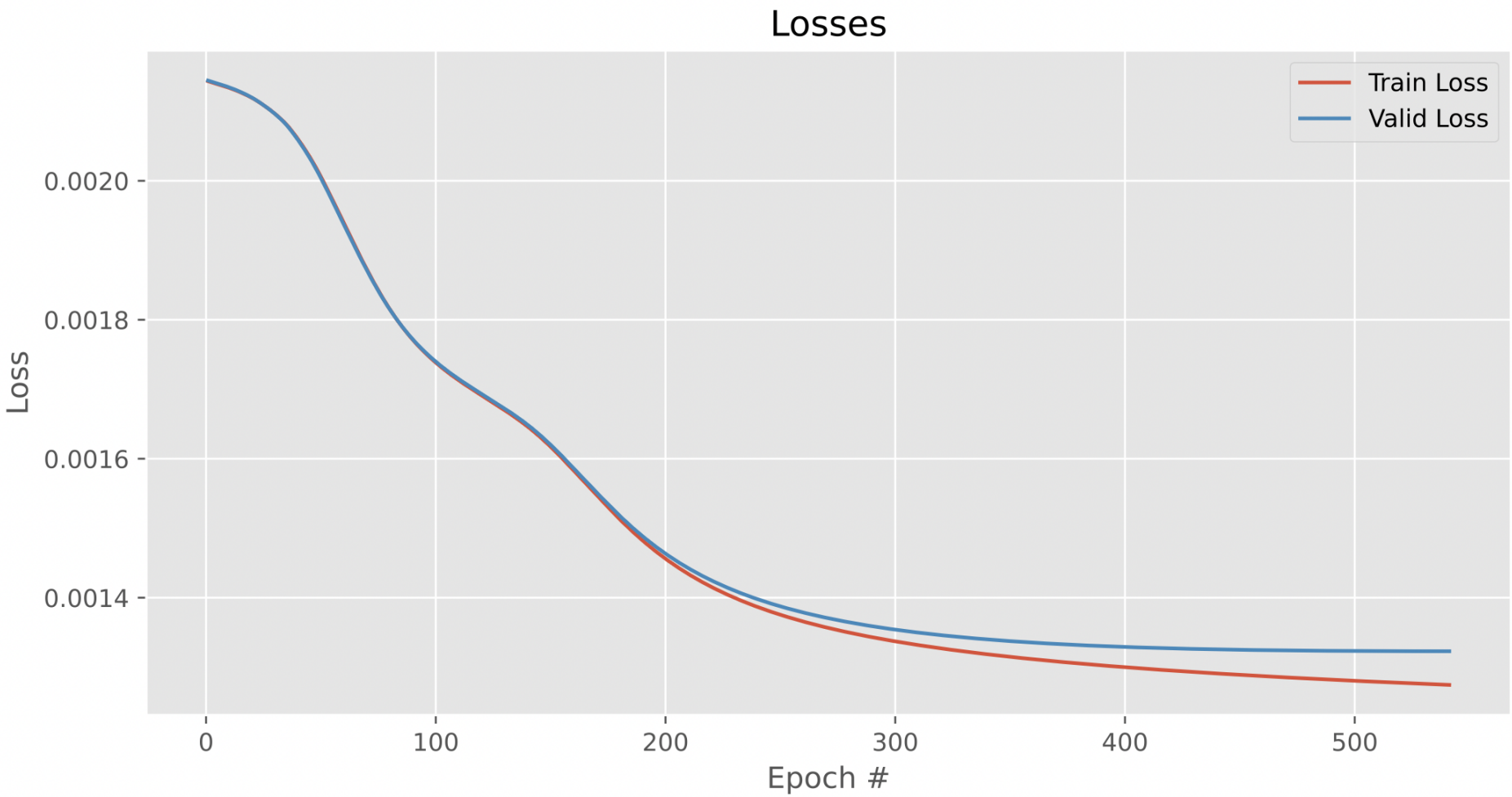


Figure: Training and Validation Losses plot

Discussion

Our model using bag of words features yield an accuracy of 83% for the two classes and F1-score of 82% for Class 0 and 85% for class 1 during evaluation with $Max\_len = 128$. When the $Max\_len = 20$ the accuracy is 86% with F1-score of 86% for the two class, the performance of the model is decreasing as the Max_len is increasing. Our model perform well when we use new data for testing.

Number	Review	Predicted	Actual
1	You are Nutzzz? I am not going to pay you Yearly for this. It is simply Not worth it.	😡	😡
2	UPDATED: I used to have this rated 4 stars until the most recent update which has issues with website addresses. Now every time you go into a note with a web address it doubles it each time you write anything and exit. PREVIOUS 4 STAR REVIEW: I was an hugh supporter of Listmaster, but they went out of business. This app does even more, except I would like to see some type of an import / export feature using csv files (for file backup).	😡	😡
3	Office 365 integration doesn't work, so it's basically useless as a business calendar. As a private calendar I don't see any big improvement over Google calendar,but it's not worse either.	😡	😡
4	Awesome app, the only missing thing is that you can only set a date, I would like to have buttons to remind me later on the day. Or on the next day, or next week. UPDATE: No what I meant, is that I want to say remind me today and not remind me at date today at specific time. UPDATE: I would then receive the notification, at a specific time 5 hours later or something that can be altered in the settings. And also I don't want to put in a date! Look at the app TickTick has exactly what I want. :)	😡	😡
5	Horrible update	😡	😡

Figure: Testing of the model using bag of words

Discussion

Our model using tf-idf have an accuracy of 83% for the two classes and F1-score of 81% for Class 0 and 84% for class 1 during evaluation.When the $Max\_len = 20$ the accuracy is 86% with F1-score of 86% for the two class, the performance of the model is decreasing as the *Max_len* is increasing.

Number	Review	Predicted	Actual
1	Very easy to use and really clear. The only thing I miss in this app is seeing the completed tasks. Besides that it's been a great app that helped my productivity.	😄	😄
2	I installed the app because it had all the features I used in my paper planner. Once I installed the app and began using the widget calendar, the numbers and saved events disappeared when I powered my Note 4 off and turned it back on.	😡	😡
3	Very very bad	😡	😡
4	i hate it😠	😡	😡
5	Really helps productivity, the email prompts and alerts are a little annoying but really helpful application which keeps my life organised love the link to Google calendar	😄	😄

Figure: Testing of the model using TF-IDF

In general, the two models (FCNN with bag of word and tf-idf) have almost the same and decent accuracy. However their performance decrease when the sentence become very long.

Discussion

- Adavantage

Fully connected layers can speed up learning and inference of the networks and make the learning problem much easier because it consist of many layers which reduces the parameters.

- Disadvantages

The number of parameters of the network are huge, therefore making the model computational expensive. The network requires a large storage space and it is difficult to optimize because the number of parameters increase with the input size.

Conclusion

To summarize: the right feature scaling can be helpful for classification.Thus, it improves the training.Linear models are the simplest models to understand, yet it still takes very careful experimentation methodology and a lot of deep mathematical knowledge to separate the theoretical and practical impacts. This might be impossible with more complicated models. Even if FCNN have decent accuracy for a certain amount of dataset for text classification, it is problematic when it comes to long sentence and large vocabulary which will lead to a big size of input features. These problem can be solve by using models with capabilities of dealing with long-term dependency data with variable input size like Long-Short Term Memory (LSTM) which is a special type of Recurrent Neural Networks (RNN).