Name : RESHMA MANOJ KUMAR
CWID : 20007266
Course : CS 541 – Artificial Intelligence
Email : rmanojku@stevens.edu

Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification

FINAL REPORT

Guided by : Prof. Abdul Rafae Khan

Sentimental LIAR: Extended Corpus and Deep Learning Models for Fake Claim Classification

Deliverable - 4

Abstract

The widespread use of social media in our daily lives and society has resulted in faster and easier access to knowledge than at any other time in human history. However, because social media platforms are fundamentally unsupervised, it has become simpler to propagate misleading information and fake news. Furthermore, the enormous volume and velocity of information flow in such platforms makes human information dissemination supervision and control impossible. This research proposes a unique deep learning strategy for automatic detection of misleading short-text claims on social media to address this issue. We begin by introducing Sentimental LIAR, which adds features based on sentiment and emotion analysis to the LIAR dataset of brief assertions. In addition, we present a new deep learning architecture based on the BERT-Base language model for determining whether claims are legitimate or not. Our findings show that the proposed architecture, when trained on Sentimental LIAR, can reach an accuracy of 70%, which is a 30% improvement above previously reported LIAR benchmark values.

Introduction

The percolation of social media throughout the world has facilitated unprecedented ease of access to the flow of in- formation. The rise of the internet and its availability have also enabled every user to to not only consume, but also contribute to the information flow. However, the benefits of such ecosystems come at the cost of mistrust in the veracity of information. In recent years, the social media scene has witnessed the proliferation of false information campaigns, in which ordinary users are intentionally or otherwise both consuming false news and also spreading it among their communities. This phenomenon is commonly referred to as fake news, broadly defined as broadcasting of information that is inten- tionally and verifiably false. The rise of fake news and its societal impact has been studied in the context of numerous recent events, such as the Brexit referendum and the 2016 US presidential elections. Fake news has thus proven to be a major threat to democracy, journalism, and freedom of expression. The exposure of users to fake news has been shown to have numerous deleterious effects, instances of which include inducing attitudes of inefficacy, alienation, trusting in false propaganda, cynicism toward certain political candidates and communities, that can at times give rise to the violent events.
In this project , Sentimental LIAR is introduced which extends the LIAR dataset by including new features based on the sentiment and emotion analysis of claims. The extended dataset also proposes a modified encoding of textual attributes to mitigate unintended bias in modeling. Furthermore, novel deep learning architecture is proposed based on the BERT- Base language model for the classification of claims as genuine or fake. The results demonstrate that the proposed architecture trained on Sentimental LIAR can achieve an accuracy of 70%, which is an improvement of 30% over previously reported results for the LIAR benchmark.

Tools and Technologies :

IDE used to run the code : Visual Studio Code – 1.63.0 (ipykernel)

Import using pandas, numpy.

DATA :

Dataset : LIAR dataset

Fake claim detection in short text has yielded a number of significant open-source datasets, some of the most notable of which are enumerated as follows:

1) FEVER [14]: Fact Extraction and VERification (FEVER) is a short claim dataset of 185,445 claims. This dataset is annotated with three labels: *Supported, Refuted,* and *Not Enough Info*. The claims are curated from Wikipedia. FEVER was constructed in two stages, the first one is Claim Generation, where extracted information Wikipedia is con- verted into claims. The second stage is Claim Labeling, where each claim is labeled as 'supported' or 'refuted' according to Wikipedia. In cases where information were insufficient for determination, a label of *not enough information* is assigned.

2) PHEME [12]: is a dataset of 330 rumor threads com- posed of 4843 tweets associated with 9 newsworthy events. The annotators of this dataset were journalists who tracked the events in real time. Each entry in PHEME is labeled as either true or false.

3) LIAR [7]: is a publicly available short statement dataset that is derived from Politifact.com. Each of the 12,836 statements in LIAR is annotated based on data available on Politifact with one of the following six labels: pants-fire, false, barely true, half-true, mostly-true, and true.

The dataset contains the text of a claim, as well as relevant meta-data, structured as follows : ID, LABEL, STATEMENT, SUBJECT, SPEAKER, SPEAKER JOB, STATE INFO, PARTY AFFILIATION, BARELY TRUE COUNTS, FALSE COUNTS, HALF TRUE COUNTS, MOSTLY TRUE COUNTS, PANTS ON FIRE COUNTS, and CONTEXT.

Authors : Bibek Upadhayay, Vahid Behzadan.

Data is preprocessed.

CODE :

```
train_size = 1
train_dataset=df.sample(frac=train_size,random_state=200).reset_index
(drop=True)
test_dataset=df_test.sample(frac=train_size,random_state=200).reset_index
(drop=True)
valid_dataset=df_valid.sample(frac=1,random_state=200).reset_index
(drop=True)
```

OUTPUT :

```
FULL Dataset: (10232, 33)
TRAIN Dataset: (10232, 33)
TEST Dataset: (1264, 32)
VALID Dataset: (1280, 33)
```

METHODOLOGY :

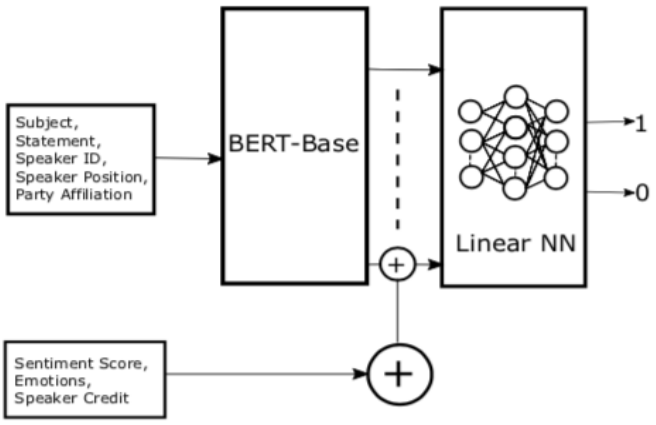
Fake claims are often written in a style of exaggerated expressions and strong emotions. Style-based classification studies aim to assess news intention. The fake claims are written with an
intention to convince the audiences to read and trust the claims, for which fake claims are written with different styles and different sentiments and emotions. With the aim of developing a computationally feasible model for fake claim detection, we propose deep neural network architectures based on BERT- Base to analyze the deception in short-text claims. Our proposed models learn to detect deception based on attribute- based language features, such as sentiments and structure- based language features. In our approach, we rely on the representation learning capabilities of transformers to extract features from statements.
We also propose an extension to the LIAR dataset that includes additional features based on the sentiment and emotion analysis of the claim.

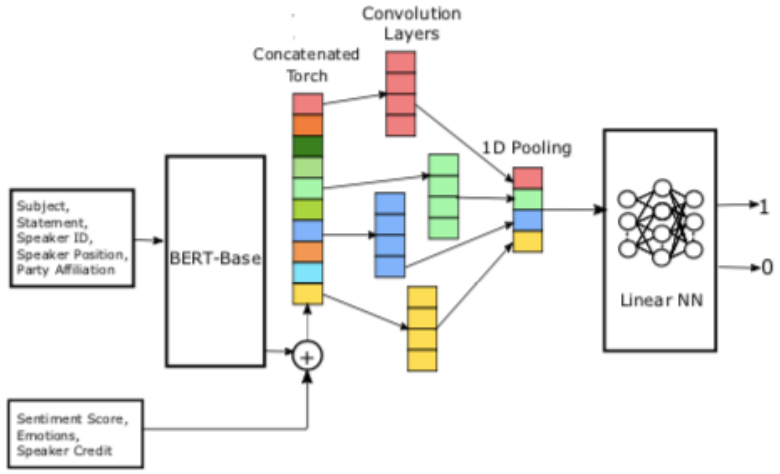
EXPERIMENTAL RESULTS :

The dataset was split into 80% train set, 10% valid set, and 10% test set. The train batch size a nd the test batch size were set to 8, with a learning rate of 1e-05. The BERT-Base was configured with a dropout of 0.3 and model used Sigmoid as its activation function. The loss function used in both architectures was the binary cross entropy loss optimized using the Adam optimizer.

Convolutional Neural Networks (CNNs) are generally known for their applications in image processing, where CNNs can learn representations of localized features from raw input while preserving the relationship between them. However, this capability of CNNs is not constrained to computer vision, and can also be adopted for NLP tasks.

MODEL :





Below demonstrates a sample record in Sentimental LIAR for a short claim in the LIAR dataset :

```
statement="McCain opposed a requirement that the government
buy American-made motorcycles. And he said all buy-American
provisions were quote ’disgraceful.’ "
subject: federal-budget
speaker id: 2
speaker job: President
state info: Illinois
party affiliation: democrat
sentiment: NEGATIVE
anger: 0.1353
disgust: 0.8253
sad: 0.1419
fear: 0.0157
joy: 0.0236
barely true counts: 70
false counts: 71
half true counts: 160
mostly true counts: 163
pants on fire counts: 9
SEN sentiment score: -0.7
```

TEXT	statement	McCain opposed a requirement that the government buy American -made motorcycles. And he said all buy-American provisions were quote 'disgraceful.'
	subject	federal-budget
	speaker_id	_2_
	speaker_job	President
	state_info	Illinois
	party_affiliation	democrat
	sentiment	NEGATIVE
EMO	anger	0.1353
	disgust	0.8253
	sad	0.1419
	fear	0.0157
	joy	0.0236
SPC	barely_true_counts	70
	false_counts	71
	half_true_counts	160
	mostly_true_counts	163
	pants_on_fire_counts	9
SEN	sentiment_score	-0.7

Experiment

This includes the training details (epochs, learning rate, dropout etc.)

```
Epoch: 0  is Started:
   Epoch: 0 Train loss is :0.6337657139998474
   Epoch 0 took: 0:06:17

Epoch: 0 - Accuracy on Testing Data Score = 0.6693037974683544
Epoch: 0 - F1 Score on Testing Data (Micro) = 0.6708860759493671
Epoch: 0 - F1 Score on Testing Data (Macro) = 0.6011597099211216

         Epoch 0 : Train Loss (Training Data):0.6337657139998474, Validatio
n Loss (Testing Data): 0.5962660192693554
_________________________________________________

         Epoch: 1  is Started:
   Epoch: 1 Train loss is :0.5931087791849804
   Epoch 1 took: 0:06:16

Epoch: 1 - Accuracy on Testing Data Score = 0.6914556962025317
Epoch: 1 - F1 Score on Testing Data (Micro) = 0.6958579881656805
Epoch: 1 - F1 Score on Testing Data (Macro) = 0.6538011695906432

         Epoch 1 : Train Loss (Training Data):0.5931087791849804, Validatio
n Loss (Testing Data): 0.5924532325387751
_________________________________________________
```

Experimental Models :

I) *Experiments with BERT-Base + Feed-Forward Neural Network*

The first method was used in three experiments with BERT- Base + Feed-Forward NN. The model was composed of BERT- Base layers, a dropout layer, and one feed-forward hidden 1 layer. The TEXT and EMO attributes were first fed into model and the output of BERT-Base (BB OP) was fed into the 0 feedforward component. This model achieved the accuracy score of 64.92% with a F1 Score of 0.6105. The input to BERT-Base was then extended by adding more meta data: TEXT+EMO+SPC and TEXT+EMO+SPC+SEN, yielding an accuracy of 67% and F1 Score of 0.40.

S.N.	Experiment	Accuracy	F1 Score Macro
1.	TEXT → [BB], BB_OP → [NN]	0.6882	0.5842
2.	TEXT+EMO → [BB], BB_OP → [NN]	0.6773	0.6352
3.	TEXT+EMO+ SPC → [BB], BB_OP → [NN]	0.6720	0.4021
4.	TEXT+EMO+ SPC+SEN → [BB], BB_OP → [NN]	0.6734	0.4097
5.	TEXT → [BB], BB_OP+EMP+ SPC+SEN → [NN]	0.6937	0.57234

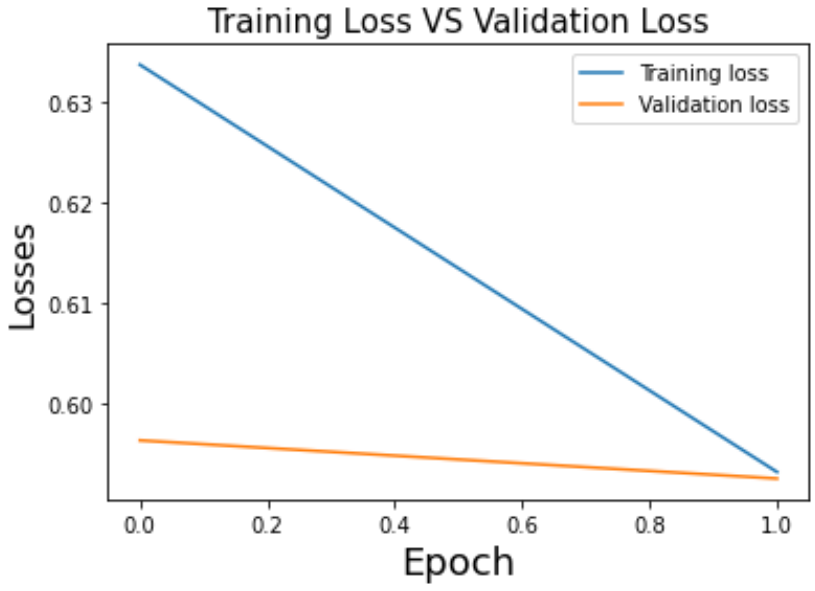
II) *Model 2: BERT-Base with CNN :*

In second model, a CNN component was appended to BERT-Base with 2 1D convolution layers. The first layer input channel size was 1 and output channel size was 50, and the second layer input channel size was 50 and output channel size was 100. The kernel size for both layers was 20 with stride of 1. The 1D max-pooling of size one was used in all the experiments. We started with the first variant, where the TEXT and other attributes were fed directly into BERT-Base, the output (BB OP) of which was passed into the CNN. In the first experiment, only TEXT was given to BERT-Base, yielding an accuracy score of 68.82% with a F1 Score of 0.5308.

S.N.	Experiment	Accuracy	F1 Score Macro
1.	TEXT → [BB], BB_OP → [CNN]	0.6882	0.5308
2.	TEXT+EMO + SPC → [BB], BB_OP → [CNN]	0.5546	0.55641
3.	TEXT → [BB], BB_OP+ EMO → [CNN]	0.6554	0.608
4.	TEXT+SPC → [BB], BB_OP+ EMO → [CNN]	0.6890	0.6542
5.	TEXT → [BB], BB_OP+EMO +SPC → [CNN]	0.7000	0.6370
6.	TXT → [BB], BB_OP+EMO+ SPC+SEN → [CNN]	0.6992	0.6430

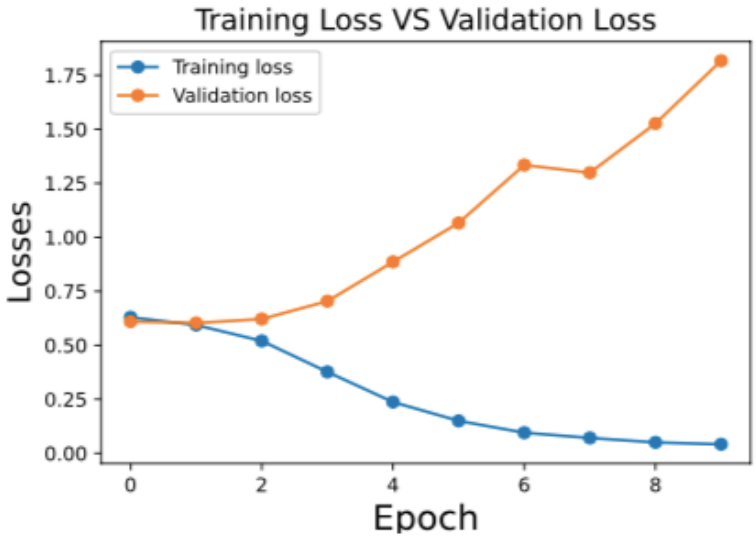
Result

The plot between Losses and Epoch.

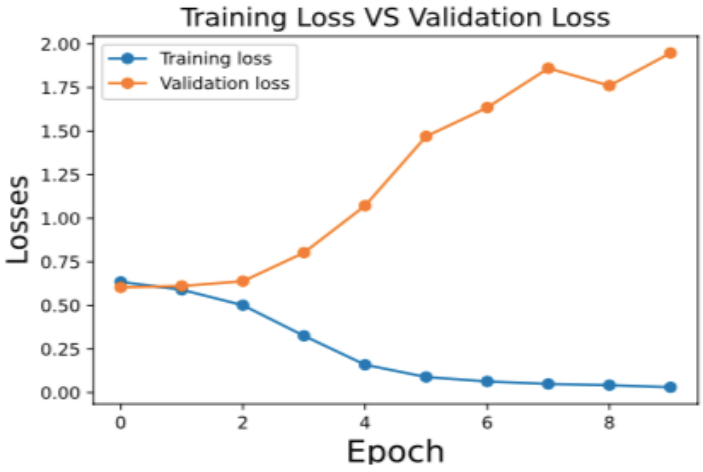



```
Epoch:  1, Accuracy Score on validation data = 0.69921875
Epoch:  1, F1 Score on Validation Data (Macro) = 0.6430236972753443
```

The training loss VS validation Loss graphs for BERT-Base + feedforward NN



The training loss VS validation Loss graphs for BERT-Base + CNN



Problems / Issues

Issues were faced when deploying certain packages and importing packages. It threw various exceptions.

Conclusion :

Sentimental LIAR was introduced as an extension of the LIAR dataset, and fresh model arch itectures based on BERT-Base for fraudulent claim detection in brief text were proposed. The proposed designs add (1) a feedforward neural network or (2) a CNN to BERT-Base.
Using IBM NLP API, the LIAR dataset was expanded to include the emotions rage, sadness, fear, anger, and disgust, as well as a sentiment score. The experiments performed with BERT -Base + feedforward NN, the accuracy ranged from 68.8% to 69% within the five experiments.
These experiments were performed by changing the input structure in the first three experi ments and by changing the hidden layers in the latter two experiments. slight improvement of 1% was observed in the accuracy and no improvements in the F1 Score. This suggests that the model may need to be revised to handle the complexity of the input data.

The experiments were performed with BERT-Base + CNN, the accuracy ranged from 68.82% to 70% within six exper- iments, and also major improvements were observed in the F1 Scor e (0.5308 to 0.6430). The best performing model is found to be one where the text attribute is fed directly into BERT-Base, and the output of BERT-Base is concatenated with the emotions , speaker's credit and sentiments before being passed to the CNN.

Furthermore, the results further verify that fake claims can be detected in short-text according to exaggerated expressions and strong emotions demonstrated in the text. The proposed architecture also sets a new state-of-the-art benchmark for fake claim classification on the LIAR dataset with an accuracy of 70%.