Yu Sze Yan Charissa a1805428@student.adelaide.edu.au

November 2021

Combating Modern Slavery with NLP

Research on Automatic Document Assessment

Research Project B

Presentation

Part 1

Introduction

1 Introduction

- This research project is a volunteering work for the non-profit org The Future Society
- Focuses on accelerating the eradication of modern slavery by applying NLP model in benchmarking modern slavery reports
- With reference to paper [1], we aim to develop a language model to classify whether companies complies with the ethical metrics
- Challenge of this project lies in the length of documents, each report have words range from 1000 to 10,000

References:
[1]‘Using augmented intelligence in accelerating the eradication of modern slavery’, by Adriana-Eufrosina Bora

Part 2

Literature Review

2 Long Document Classification related works

- Transformers suffer from major issue of applicability in the classification of long sequences.
- Paper [2] proposed Model Recurrence over BERT (RoBERT), to address this limitation, in applying to classifying transcripts of human call conversations
- Input is segmented into smaller chunks and fed into the base model. Each output is propagated through a single recurrent layer followed by a softmax activation.

References:
[2]BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, by Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
[3]Hierarchical Transformer for Long Document Classification, Raghavendra Pappagari1, Piotr Z elasko2, JesusVillalba1, Yishay Carmiel2, and Najim Dehak

Part 3

Dataset and Experiment Set up

3 Data-set & Experiment Set up

- We take metrics ‘whistleblower’ as the experimenting metric.
- It determines whether businesses have provided a mechanism, e.g. a hotline, for anyone in their operations may report incidents of slavery to a focal point; and whistle-blowers are protected
- It consists of 1343 business reports, with average document length over 1500 words.

3 Experiment Set up Summary

To tackle the challenge of long document, we break down classification task into 2 steps:

- 1) selecting the key sentences, encoding them into sentence vectors, concatenating sentence vectors to form document representation and
- 2) classifying the document representation

	BERT	TF-IDF
Document representation	• Direct keyword sentence extraction or • Weighted sentence attention system	• Direct keyword sentence extraction
Classification	• Neural Network (NN)	• Supported Vector Classifier (SVC) or • Neural Network (NN)

Part 4

Experimenting Models (1) BERT

4 BERT

4.1. Baseline model

4.2. Direct keyword sentence extraction without keyword attention mechanism

4.3 Direct keyword sentence extraction with keyword attention mechanism

4.4.1 One-step weighted sentence attention model

4.4.2 Two-step weighted sentence attention model

4.5 Data augmentation

4 BERT

	Baseline 4.1	Direct keyword sentence extraction without keyword attention mechanism 4.2	Direct keyword sentence extraction with keyword attention mechanism 4.3	One-step weighted sentence attention model 4.4.1	Two-step weighted sentence attention model 4.4.2	Data augmentation 4.5
Sentence attention (1): Direct keyword sentence extraction		V	V			
Sentence attention (2): Soft sentence attention				V	V	V
Word attention : Keyword-string paring			V	V	V	V

4.4.1 One-step weighted sentence attention model

- In the key sentence selection, we applied an attention mechanism to sentences.
- Each sentences x_i in the document d is represented in BERT, output as zi. In f(zi), we find the weight w_i applied to the sentence through training, where higher value is applied to more important sentence and lower value for less important one. Therefore it is a soft-cut mechanism for selecting key sentences.
- The attention mechanism is further intensified by adding lambda times I_i in the f(zi).
- After normalisation, weight of sentences does not contain any useful information for the classifying task would become zero. By multiplying 0 with the sentence vector, the sentence is removed from the classifying text.
- Lastly, a neural network is used to perform classification of the final document representation h.

$$d = \{x_1 \ldots. x_i\}$$

$$z_i = \text{BERT } (x_i)$$

$$f(z_i) = w^T z_i + b + \lambda I(x_i, \{\text{keywords}\})$$

$$I(x_i, \{\text{keywords})\} = \begin{cases} 1, & \text{if x has keywords} \\ 0, & \text{otherwise} \end{cases}$$

$$\alpha_i = \exp(f(z_i)) / \sum_j \exp(f(z_i))$$

$$h = \sum_i \alpha_i z_i$$

$$cls(h)$$

4.4.2 Two-step weighted sentence attention model

- Unlike model 4.4.1, we omit lambda by having a two step training, due to the possibility that adding lambda in f(xi) weakens the training of weight w_i, so the model would solely depends on lambda for selection of key sentences.

- In the first step , the weight is trained to predict whether the sentence is a keyword sentence by minimizing the KL divergence between f(zi) and l_i. It is guided to produce alpha_i = 1 for keyword sentence and 0 for non-keyword sentence after normalization. In step two, the weight is trained to give importance to sentence x_i without any guidance.

- Lastly, neural network is used to perform classification of the final document representation h.

$$d = \{x_1 \ldots . x_i\}$$

$$z_i = \text{BERT}\ (x_i)$$

$$f(z_i) = w^T z_i + b$$

$$\alpha_i = \exp(f(z_i)) / \sum_j \exp(f(z_i))$$

$$h = \sum_i \alpha_i z_i$$

$$cls(h)$$

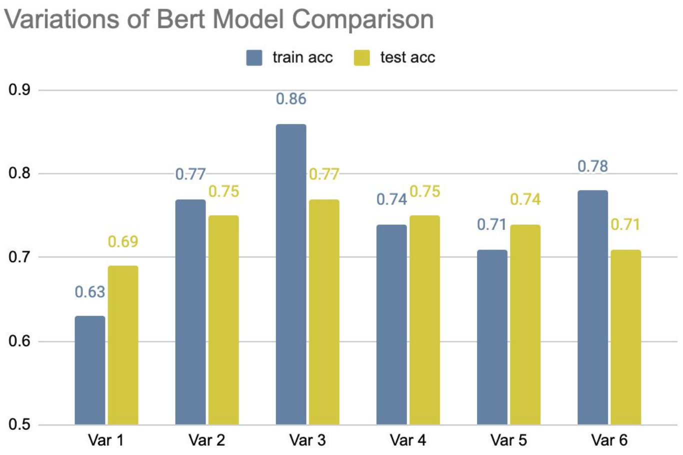
4.6 BERT - Result and Analysis

We evaluate the utility of keyword attention mechanism, (var.2 & var3)., model is improved by 2%.

We evaluate the utility of sentence attention mechanism (var.3., var.4)., Testing accuracy dropped by 2%, shows that adding more information out of the keyword sentences in the soft sentence attention mechanism does not improve model.

Nonetheless, the comparable accuracy between direct key-sentence extraction model and one-step sentence attention model (var3 & var 4) shows that soft sentence attention mechanism is effective in selecting the key sentence for classification.

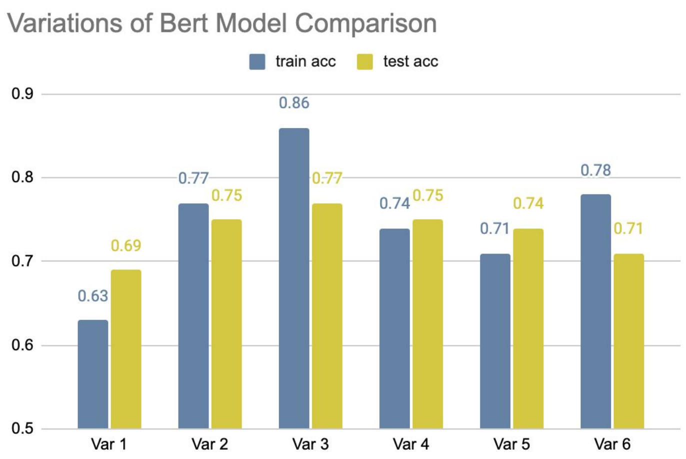
Comparing the model with and without applying data augmentation (var4 , var.6). Testing accuracy has dropped by 4%, adding simulated samples does not reduce overfitting issue.



4.6 BERT - Result and Analysis

In overall, direct key sentence extraction with keyword attention model (var.3) is the best performing variations with 77% testing accuracy; compared to baseline no attention mechanism, model improved 8%.

Since key sentence selection mechanism is effective, the average-performance of BERT model is attributed to the classifying step, where the representation of key sentence plays a crucial part.



Part 5

Experimenting Models (2)
TF-IDF

5.1 TF-IDF

In TF-IDF model, keyword sentences are extracted and concatenated; generated into document representation in different settings in terms of composition and n-gram:

(1) Composition: the full concatenated sentence is encoded in TF-IDF model or keywords only in the concatenated sentence is encoded.

(2) N-gram: Uni-gram, bi-gram or both uni-gram and bi-gram

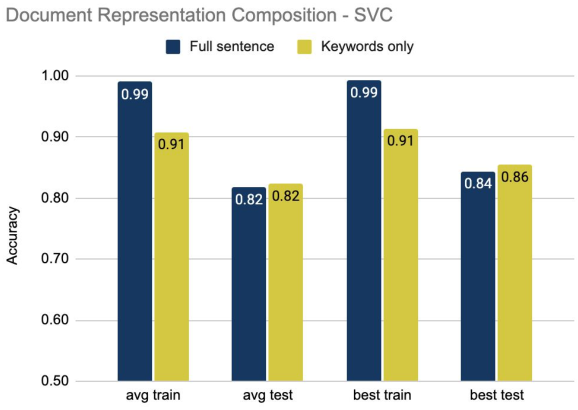
Document representation is put into 2 classifiers - Supported Vector classifier (SVC) and Neural Network (NN).

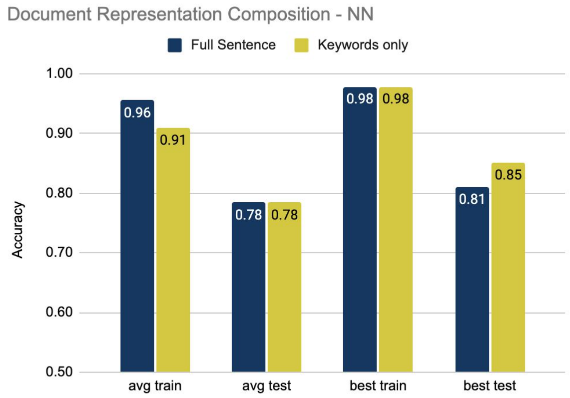
Model is run with all variation combinations and 10 runs of random train-test split.

	TF-IDF
Composition	• Full concatenated sentence or • Extracting keywords only from the concatenated sentence
N-gram	• Uni-gram • Bi-gram • Uni-gram & Bi-gram
Classifier	• SVC • NN

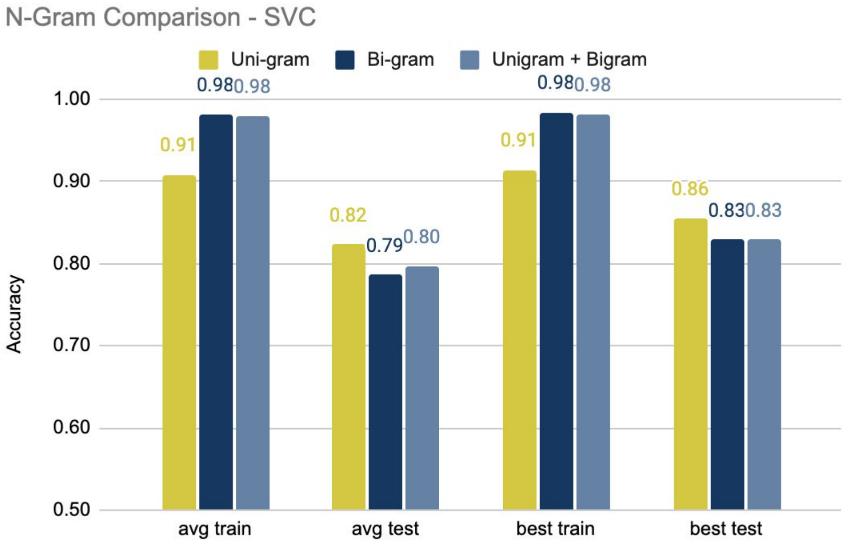
5.2 Composition of Input

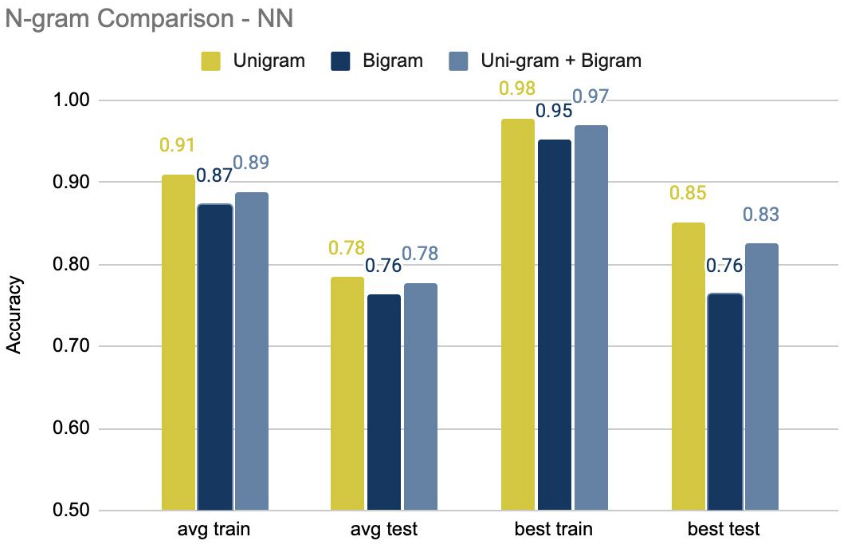
Keywords-only vector contributes to a better result in both classifiers.





5.3 N-Gram


N-Gram Comparison - SVC
Uni-gram
Bi-gram
Unigram + Bigram
Accuracy
1.00
0.90
0.80
0.70
0.60
0.50
0.91
0.98
0.98
0.82
0.79
0.80
0.91
0.98
0.98
0.86
0.83
0.83
avg train
avg test
best train
best test


N-gram Comparison - NN
Unigram
Bigram
Uni-gram + Bigram
Accuracy
1.00
0.90
0.80
0.70
0.60
0.50
0.91
0.87
0.89
0.78
0.76
0.78
0.98
0.95
0.97
0.85
0.76
0.83
avg train
avg test
best train
best test

5.3 N-Gram

Uni-gram outperforms bi-gram and combination of the two in both classifiers.

N-gram = 1 allows flexibility of variation in keywords

From the samples, ’senior manager’, ’managing director’ , ’line manager’, are the keywords with personnel meaning which its occurrence is important for classification.

As the word ‘manager’ is standardised to ‘manag’, with uni-gram, it give the keyword ’manager’ a certain degree of variation flexibility for classification.

Positive sample 47: *“If a hotel employee witnesses an indicator leading them to sus- pect human trafficking or other types of modern slavery, they must inform the **head of department** and **senior manager** on duty immediately, in order to submit an internal report.”*

stemmed keywords: ‘head’, ‘depart’, **‘senior’, ‘manag’**

Positive sample 726: “If employees see anything that goes against our Values, breaks the law, breaches our regulations or policies, or simply feels wrong,they are encouraged to speak to their ***line managers***, ***HR*** or ***Compliance***.”

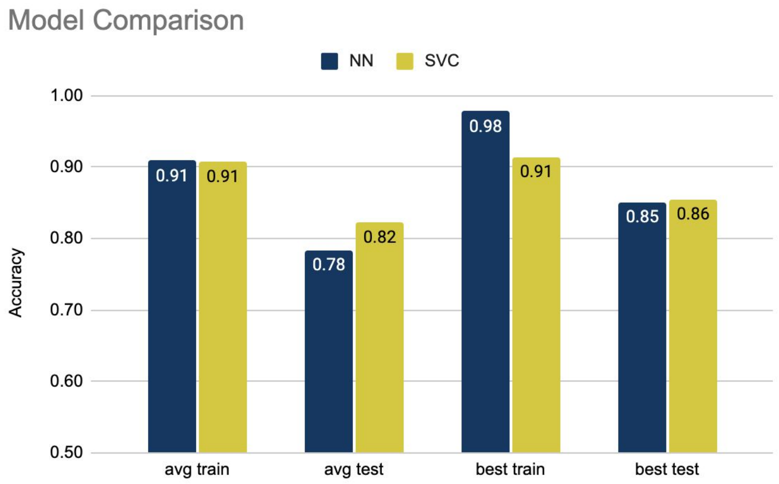
stemmed keywords: **‘line’, ‘manag’**, ‘HR’, ‘Compliance’

Positive sample 1300: “We have whistle-blowing procedures in place which allow our team to raise concerns of any nature internally to our ***Managing Director*** , or externally via a dedicated neutral phone service.”

stemmed keywords: **‘manag’, ‘director’**

5.4 Classifier

Results shows SVC outperforms NN model by 1% in best performing testing accuracy and by 4% in average 10- runs testing accuracy.



Part 6

Analysis of Models

6.1 BERT & TF-IDF model

	BERT	TF-IDF
Key sentence selection	Direct keyword sentence extraction	Direct keyword sentence extraction
Key sentence representation	Keyword-string pairing attention	Keywords only with Uni-gram
Classification of selected key sentence	Fully-connected neural network	SVC
Training Acc	0.86	0.91
Testing Acc	0.77	0.86

6.1 BERT & TF-IDF model

From analysis section 4.6, the average-performance of BERT lies in the sentence vector representation.

We evaluate the distinguishability of BERT and TF-IDF sentence vector by calculating their **cosine similarity** with data samples.

In BERT, adding sentence pair entailment is not generating a sentence representation strong enough paying attention to a particular few keywords.

TF-IDF, keywords are embedded individually in the sentence vector.

Negative sample 89:

’We encourage all our employees to report any concerns either internally or via our Whistleblowing Policy.’

Positive sample 1000:

’Whistleblowing Policy - we encourage all employees, customers and suppliers to report any suspicion of slavery or human trafficking ***without fear of retaliation****. ’*

model	Bert 0	Bert 1	Tf-Idf 0	Tf-Idf 3
pair 1	0.997	0.985	0.078	0.197

Cosine Similarity of sentence vector from BERT & Tf-idf

6.2 Data-set & model

6.2.1 Limited vocabulary variation

Vocabulary used in official documents is limited and standard, makes the model training not able to benefit from adding more augmented data, to tackle overfitting issue caused by small sample size.

BERT pre-trained model is useful for document classification that there is lots of variation in vocabulary, especially those vocabulary are not seen in training stage yet have similar semantic meaning.

Limited vocabulary feature gives advantage in applying TF-IDF sentence representation, as the range of vocabulary problem is avoided.

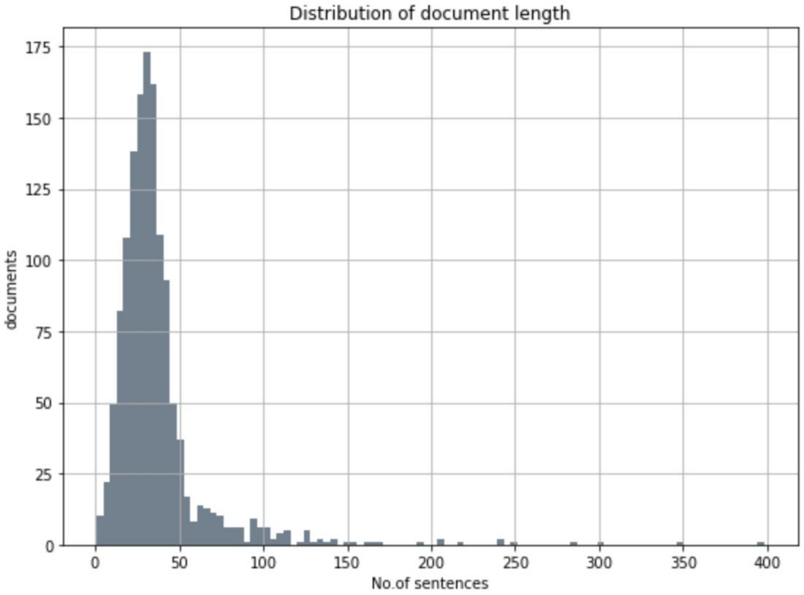
6.2 Data-set & model

6.2.2 Information concentrated on few sentences

Number of sentences ranges from 5 to 395, yet classification information concentrating on only 1-3 sentences.

Soft sentence attention mechanism in BERT is more applicable with classification information spreading across many sentences with varying importance, therefore hard sentence attention is more applicable.

Information concentrating on few sentences also implies that the classification task requires less importance in capturing the sequential relationship among sentences, hence RoBERT model is also not applicable.



6.2 Data-set & model

6.2.3 High recall rate of samples having keywords

High percentage (90%) of samples having keywords enables direct keyword sentence extraction to be valid, and making it more applicable than soft sentence attention mechanism.

Combined with the feature of keywords being standard and limited, two features making this classification task a keyword combination optimization problem, rather than a semantic problem.

Hence enable TF- IDF being a model retains no information on grammar, nor word order outperforms the strong-semantic-capability BERT model in solving a keyword combination optimization problem.

	Positive	Negative
Keyword	0.51	0.39
No Keyword	0.01	0.09

6.2 Data-set & model

6.2.4 High similarity text differentiable by keywords

The classified text of this data-set is highly similar and the differentiable depends on the occurrence of few keywords.

Illustrated in 6.1 cosine similarity example.

Part 7

Conclusion

7 Conclusion

- Sequence models like RoBERT is not applicable when sequence relationship between sentences is not important to perform classification. We proposed the sentence attention mechanism model.

- Breaking long document assessment task into key sentence selection and classification of key sentence, we effectively implemented a soft key sentence attention mechanism in BERT model, a more robust system adding flexibility to handle documents with sentence does not contain keyword yet is important for classification.

- While on the classifying step, with the nature of limited vocabulary variation and being differentiated by occurrence of keywords in this data-set, the TF-IDF sentence vector representation combined with SVC, is able to perform long document assessment task of bench-marking Modern Slavery Statements and achieving testing accuracy of 86%.

- TF-IDF sentence vector representation in combined with SVC, could be viewed as a weighted keyword combination optimization algorithm, well-suited for classifying key-word focus documents with standard and narrow vocabulary, such as official statements or legal documents.

Thank You!

Please free free to ask any questions.

Supervisor: Dr. Lingqiao Liu

Student Name : Yu Sze Yan Charissa

Student no: a1805428