e-ISSN: 2540-9824, p-ISSN: 2540-9433

+62-341-577911 jitecs@ub.ac.id (mailto:jitecs@ub.ac.id)

(https://jitecs.ub.ac.id/index.php/jitecs/index)

Vol. 5 No. 3: Desember 2020

The articles in this issue (10 original research articles) were authored/co-authored by 27 authors from 2 countries (Indonesia, Nigeria).

DOI: https://doi.org/10.25126/jitecs.202053 (https://doi.org/10.25126/jitecs.202053)

PUBLISHED: 2020-12-31

ARTICLES

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/183/129)

RESTful API Implementation in Making a Master Data Planogram Using the Flask Framework (Case Study: PT Sumber Alfaria Trijaya, Tbk) (https://jitecs.ub.ac.id/index.php/jitecs/article/view/189)

Era Susanti, Evangs Mailoa
255-269

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/189/130)

Analysis in the Strategic Formula for Business and Information Technology Alignment of the Research and Development Planning Institution in Batu City (https://jitecs.ub.ac.id/index.php/jitecs/article/view/214)

Aulia Zahra Musthafawi, Ismiarta Aknuranda, Fatwa Ramdani
270-278

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/214/131)

Automated Features Extraction from Software Requirements Specification (SRS) Documents as The Basis of Software Product Line (SPL) Engineering (https://jitecs.ub.ac.id/index.php/jitecs/article/view/219)

M Syauqi Haris, Tri Astoto Kurniawan, Fatwa Ramdani
279-292

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/219/132)

Promoting Interest in Learning YorÃ¹bÃ¡ Language Using Mobile Game (https://jitecs.ub.ac.id/index.php/jitecs/article/view/232)

Oladosu Oladimeji, Temitope Olorunfemi, Olayanju Oladimeji
293-301

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/232/133)

Development of Big Data App for Classification based on Map Reduce of Naive Bayes with or without Web and Mobile Interface by RESTful API Using Hadoop and Spark (https://jitecs.ub.ac.id/index.php/jitecs/article/view/233)

Imam Cholissodin, Diajeng Sekar Seruni, Junda Alfiah Zulqornain, Audi Nuermey Hanafi, Afwan Ghofur, Mikhael Alexander, Muhammad Ismail Hasan
302-312

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/233/134)

Comparison of Neural Network and Recurrent Neural Network to Predict Rice Productivity in East Java (https://jitecs.ub.ac.id/index.php/jitecs/article/view/182)

Andi Hamdianah, Wayan Firdaus Mahmudy, Eko Widaryanto

313-324

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/182/136)

Utilizing Indonesian Universal Language Model Fine-tuning for Text Classification (https://jitecs.ub.ac.id/index.php/jitecs/article/view/215)

Hendra Bunyamin

325-337

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/215/137)

Mobile Application Architecture Restructuring with Microservice Approach (https://jitecs.ub.ac.id/index.php/jitecs/article/view/239)

Ardiono Roma Nugraha, Aini Suri Talita

PDF (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ARTICLE/VIEW/239/138)

ARTICLE SUBMISSION GUIDELINES

ARTICLE SUBMISSION GUIDE (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ABOUT/SUBMISSIONS#AUTHORGUIDELINES)
PUBLICATION ETHICS (HTTPS://JITECS.UB.AC.ID/INDEX.PHP/JITECS/ABOUT#PUBLICATION-ETHICS)
DOWNLOAD TEMPLATE
PUBLICATION FEE

ABOUT THIS JOURNAL

Focus and Scope (https://jitecs.ub.ac.id/index.php/jitecs/about#focusAndScope)

Publishing Frequency

Peer Review Process (https://jitecs.ub.ac.id/index.php/jitecs/about#peerReviewProcess)

Open Access Policy (https://jitecs.ub.ac.id/index.php/jitecs/about#openAccessPolicy)

Copyright Notice

Contact (https://jitecs.ub.ac.id/index.php/jitecs/about/contact)

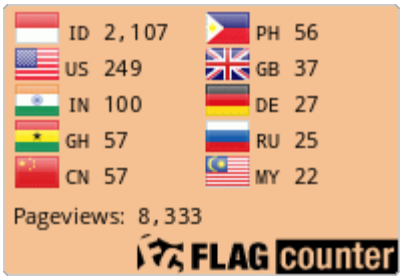
ACCREDITATION CERTIFICATE

(https://jitecs.ub.ac.id/public/site/images/sigit/jitecs-certificate.jpg)

JITeCS has been accredited by the Ministry of Research, Technology, and Higher Education

Accreditation 85/M/KPT/2020 (http://arjuna.ristekbrin.go.id/files/info/Hasil_Akreditasi_Jurnal_Nasional_Periode_1_Tahun_2020.pdf)

VISITORS



(https://info.flagcounter.com/hOGs)

CITATIONS & REFERENCE MANAGER

(https://www.mendeley.com/)

CURRENT ISSUE

ATOM 1.0 (https://jitecs.ub.ac.id/index.php/jitecs/gateway/plugin/WebFeedGatewayPlugin/atom)

RSS 2.0 (https://jitecs.ub.ac.id/index.php/jitecs/gateway/plugin/WebFeedGatewayPlugin/rss2)

RSS 1.0 (https://jitecs.ub.ac.id/index.php/jitecs/gateway/plugin/WebFeedGatewayPlugin/rss)

INFORMATION

For Readers (https://jitecs.ub.ac.id/index.php/jitecs/information/readers)

For Authors (https://jitecs.ub.ac.id/index.php/jitecs/information/authors)

For Librarians (https://jitecs.ub.ac.id/index.php/jitecs/information/librarians)

Supported by

(http://filkom.ub.ac.id)

(http://ub.ac.id)

Technical Support

(https://www.doi.org/)

(http://turnitin.com/)

(http://mendeley.com/)

FILKOM Links

Laboratories

Research Laboratory (http://filkom.ub.ac.id/legacy/unit/lab/kcv)
Learning Laboratory (https://filkom.ub.ac.id/lab-pembelajaran/)

Student

Student Executive Agency (http://bemtiik.ub.ac.id/)
Student Representative Council (http://dpmptiik.ub.ac.id/)
Student Association (http://hmif.ub.ac.id/)
Display (http://display.ub.ac.id/)
Raion (http://raioncomm.ptiik.ub.ac.id/)
K-Risma (http://krismatiik.ub.ac.id/)
Optiik (http://optiik.ub.ac.id/)
Poros (http://poros.ub.ac.id/)
Bios (http://bios.ptiik.ub.ac.id/)

FILKOM Directory

PK2maba (http://filkom.ub.ac.id/event/pk2maba)
Journal (http://jtiik.ub.ac.id)
Student Journal (http://j-ptiik.ub.ac.id)

News (https://filkom.ub.ac.id/berita/)
Map (http://filkom.ub.ac.id/legacy/info/map)

UB Directory

Lecture Blog's (http://lecture.ub.ac.id)
Staff Blog's (http://staff.ub.ac.id)
E-Complaint UB (http://e-complaint.ub.ac.id/)
Online Scholarship UB (http://beasiswa.ub.ac.id/)
Job Placement Center UB (http://jpc.ub.ac.id/)

Tweet

Tweet dari @filkomUB

FILKOM UB
@filkomUB · 23j

Sosialisasi Program MSIB & Bangkit Semester Ganjil 2023/2024.

Journals

Sort by

Impact

Search journals

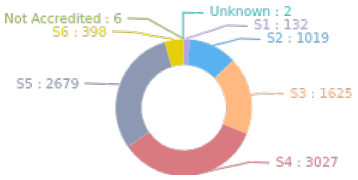
Search...

Filter

8.888
Total Journals

1.345
Total Publishers

Acceditations Level

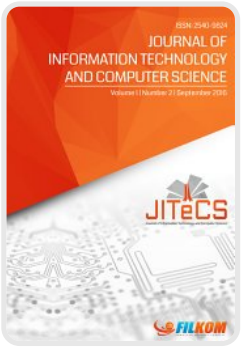


Results for *"Jitecs"*

× clear search

Previous 1 Next

Page 1 of 1 | Total Records 1



JITECS (JOURNAL OF INFORMATION TECHNOLOGY AND COMPUTER SCIENCE)

Google Scholar Website Editor URL

Fakultas Ilmu Komputer, Universitas Brawijaya

P-ISSN : 25409433 | E-ISSN : 25409824 Subject Area : Science, Engineering

S3 Accredited Garuda Indexed

1,86
Impact

9
H5-index

421
Citations 5yr

432
Citations

Previous 1 Next

Page 1 of 1 | Total Records 1

Journals

Sort by

Impact

Search journals

Filter

Journal of Information Technology and Computer Science
Volume 5, Number 3, December 2020, pp. 325-337
Journal Homepage: www.jitecs.ub.ac.id

Utilizing Indonesian Universal Language Model Fine-tuning for Text Classification

Hendra Bunyamin[1]

[1]Informatics Engineering, Maranatha Christian University,Bandung
[1]hendra.bunyamin@it.maranatha.edu

Received 28 June 2020; accepted 30 December 2020

Abstract. Inductive transfer learning technique has made a huge impact on the computer vision field. Particularly, computer vision applications including object recognition, segmentation, and classification, are seldom trained from scratch; instead, they are fine-tuned from pretrained models, which are products of learning from huge datasets. In contrast to computer vision, state-of-the-art natural language processing models are still generally trained from the ground up. Accordingly, this research attempts to investigate an adoption of the transfer learning technique for natural language processing. Specifically, we utilize a transfer learning technique named Universal Language Model Fine-tuning (ULMFiT) for doing an Indonesian news text classification task. The dataset for constructing the language model is collected from several news providers from January to December 2017 whereas the dataset employed for text classification task comes from news articles provided by the Agency for the Assessment and Application of Technology (BPPT). To examine the impact of ULMFiT, we provide a baseline that is a vanilla neural network with two hidden layers. Although the performance of ULMFiT on validation set is lower than the one of our baseline, we find that the benefits of ULMFiT for the classification task significantly reduce the overfitting, that is the difference between train and validation accuracies from 4% to nearly zero.

Keyword : Learning, model, classification, training

1 Introduction

Text classification is defined as a problem of formulating models for organizing documents into pre-defined labels [27,26,31]. Nowadays, researchers have been developing systems with new text classification techniques with goals to achieve better computational efficiency and prediction accuracy than before[45,1,36,31,2].

Recently deep learning gains increasing popularity from renowned scientists because of its wide range applications in many fields [15,43] such as object recognition in photos [24,35], speech recognition [17], pedestrian detection and image segmentation [38,12,9], traffic sign classification [8], machine translation [41,3], playing Atari video games [32], reinforcement learning for robotics [13], assisting pharmaceutical company in designing new medicine [10], searching subatomic particles [4], and automating process of parsing microscope images for building 3-D maps of human brains [23].

The state-of-the-art text classification techniques are largely based on deep learning. Basically, neural networks with multiple hidden layers are considered as deep learning as depicted in Fig. 1. What special about deep learning is its capability of learning non-linear relationship from dataset [20]. The success stories of deep learning begin in computer vision field, specifically in image classification competition known as ImageNet [14]. Deep learning also enjoys its glory in natural language processing (NLP) field, notably automatic translation system as reported in New York Times magazine articles [25].

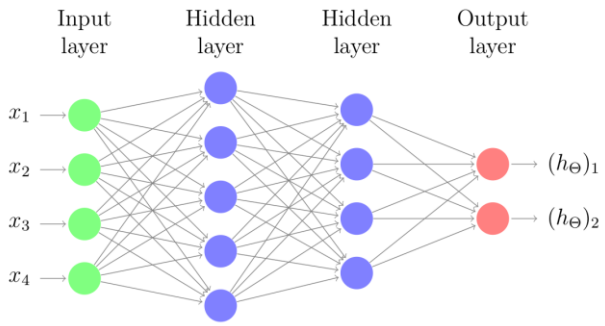


Fig. 1. A neural network with two hidden layers; the more hidden layers there are, the better the performance will be (the principle of deep learning)

In general, there are two factors that drive the success of deep learning [33]. Firstly, the computational power that has the capability to compute weights of large neural networks grows exponentially through years. Secondly, huge amount of available data which is the source of training set has made the networks learn better than before. Consequently, significant NLP progresses can be done by institutions that have facilities to collect massive datasets, give labels to them, and do data crunching for training purposes. However, these costly training time may be reduced by employing transfer learning concept [21]. In short, transfer learning allows pretrained models to be reused for solving similar tasks. Well-known examples of transfer learning are utilizing pretrained networks such as Xception [6], Inception [42], ResNet50 [16], VGG16 [39], and MobileNet [37] for image classification.

This research is related to text classification; specifically, we investigate the use of transfer learning for doing Indonesian news text classification task and overcoming the overfitting problem. Experiments are performed in order to validate whether or not transfer learning can leverage text classification algorithms to increasing their performance and mitigating the overfitting problem. Regarding the choice of transfer learning algorithms, we opt for Universal Language Model for Fine-tuning (ULMFiT) because this algorithm has been reported as one of the state-of-the-art transfer learning algorithms [21].

2 Literature Review

Prior work has documented three interconnected components which build ULMFiT such as regularization, weight-dropped LSTM, and transfer learning techniques.

2.1 Weight-dropped LSTM

The Long short-term memory (LSTM) is a popular variant of recurrent neural network [18,11] that is robust to vanishing or exploding gradients [5]. LSTM is formulated as follows:

$$i^{(t)} = \sigma(W_i x^{(t)} + U_i h^{(t-1)}) \quad (1)$$

$$f^{(t)} = \sigma(W_f x^{(t)} + U_f h^{(t-1)}) \quad (2)$$

$$o^{(t)} = \sigma(W_o x^{(t)} + U_o h^{(t-1)}) \quad (3)$$

$$\tilde{c}^{(t)} = \tanh(W_c x^{(t)} + U_c h^{(t-1)}) \quad (4)$$

$$c^{(t)} = i^{(t)} \odot \tilde{c}^{(t)} + f^{(t)} \odot \tilde{c}^{(t-1)} \quad (5)$$

$$h^{(t)} = o^{(t)} \odot \tanh(c^{(t)}) \quad (6)$$

with W_i, W_f, W_o, U_i, U_f, U_o are weight matrices, $x^{(t)}$ is input vector at timestep t, $h^{(t)}$ is hidden state at timestep t, $c^{(t)}$ is a state of memory cell, and ○ denotes an element-by-element multiplication.

Weight-dropped LSTM is a modification to standard LSTM by adding a DropConnect [44] operation on the recurrent hidden to hidden weight matrices, specifically weight matrices $\{U_i, U_f, U_o, U_c\}$ in Eq. (1), (2), (3), (4) respectively. Moreover, the use of DropConnect prevents LSTM from overfitting. The Weight-dropped LSTM is commonly called AWD-LSTM [28].

2.2 Regularization

Regularization is a common technique in machine learning used to fight overfitting problem. Overfitting occurs when a text classification model has achieved an almost perfect performance on training set (Fig. 2) whereas its performance is very bad on testing set (Fig. 3).

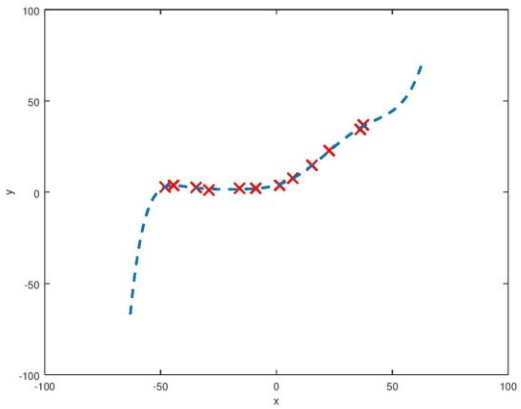


Fig. 2. Overfitting problem with *x*-axis is a feature and *y*-axis is a label

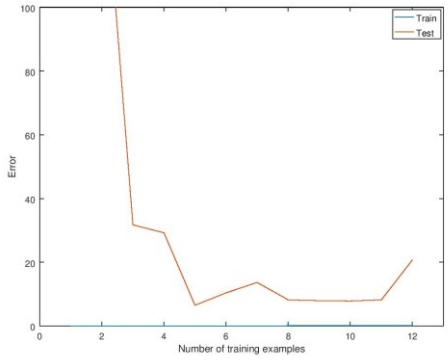


Fig. 3. Overfitting symptoms: a zero-approaching train error and a high test error

Cost function of a neural network without regularization is defined as follows:

$$\min_{\theta}\left[\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2\right] \qquad (7)$$

with $h_\theta(x^{(i)})$ is our model prediction for $x^{(i)}$, θs are parameters of the model, and $y^{(i)}$ is the ith label for the ith feature. In order to mitigate the overfitting, a constraint is added into Eq. (7);

$$\min_{\theta}\left[\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\underbrace{\lambda\sum_{j=1}^{n}\theta_j^2}_{\text{regularization}}\right] \qquad (8)$$

therefore, the cost function with regularization becomes with λ is a regularization parameter and θ_j are parameters of the model indexed by j=1, ..., n.

There are four additional regularization techniques that reduce data storage during model training and prevent overfitting of the LSTM model. Firstly, the variable regularization techniques allows an efficient data usage by randomly select a sequence length for forward and backward propagation. Secondly, the embedding dropout helps the model perform dropout on the embedding matrix at a word level; specifically, the dropout is broadcast across all word vector embeddings. The last two techniques are related to L_2-regularization. Activation regularization used to keep outputs of activation functions from significantly greater than 0, whereas temporal activation regularization prevents the model from generating great changes in the hidden state [29].

2.3 Transfer learning

Transfer learning is a technique of utilizing a pretrained network [7]. There are two techniques of employing the pretrained network that are feature extraction and fine-tuning. Specifically, Universal Language Model Fine-tuning (ULMFiT) [21] does an initial training for constructing a language model (LM) on general-domain corpus and fine-tuning the LM on a target task, for example, text classification.

Concretely, Fig. 4 shows three steps of ULM-FiT construction. Firstly, a LM

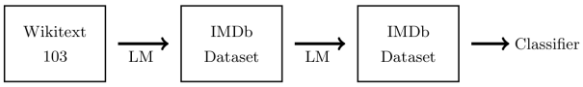


Fig. 4.A high-level overview of constructing ULM-FiT

is constructed by learning from a large dataset, Wikitext 103; owing to the movie review classification target, the LM is fine-tuned on IMDb corpus. Finally, the LM is trained to be ready as a classifier.

3. Research Methodology

This section starts with the description of our dataset. Next, we elaborate the steps of constructing the ULM-FiT. The target task in this research is to classify news articles released by the Agency for the Assessment and Application of Technology (BPPT). The dataset is available to download at `https://github.com/hbunyamin/BPPT-dataset`. Specifically, Table~\ref{tab:contoh-dokumen-berita} displays samples of the BPPT dataset.

Text	Label
Pemain timnas Jerman Lehmann kehilangan tempatnya di Arsenal dan digantikan Manuel Almunia dan Dortmund mencoba untuk melakukan kesepakatan dan membawanya ke Bundesliga.	Sports
Menurut Edwin, kenaikan harga BBM ini merupakan isu global yang terjadi hampir di semua negara bukan hanya di Indonesia. Kenaikan harga BBM tidak dapat ditunda lagi harus segera dilakukan pada awal bulan ini.	Economics
Kementerian Kesehatan Irak dipandang sebagai sebuah tempat yang aman bagi milisi Syiah, dan baik al-Zamli maupun Menteri Kesehatan Ali al-Shemari adalah anggota kelompok garis keras Syiah pimpinan Moqtada al-Sadr.	International
Salah satu pengembangan itu adalah makanan transgenik.	Science

Table 1. Four examples of news articles and labels

3.1 Preprocessing the dataset

Next, Table 2 displays the number of articles in each label. Besides BPPT dataset, our second dataset comprises a collection of several articles from news provider such as liputan6.com, kompas.com, tribunnews.com, detik.com, beritasatu.com, and tempo.co during year 2017. Particularly, total number of collected articles are 455.626. This dataset is used for training a pre-trained language model which is one of the ULMFiT model construction ingredients. A sample of the preprocessed news article just before tokenization process is shown in Table 3.

Filename	Category	# Sentences
`PANL-BPPT-ECO-ID-150Kw.txt`	Economics	6.544
`PANL-BPPT-INT-ID-150Kw.txt`	International	6.642
`PANL-BPPT-SCI-ID-100Kw.txt`	Science	6.355
`PANL-BPPT-SPO-ID-100Kw.txt`	Sports	4.483

Table 2. Filenames with labels and number of sentences

xxbos surabaya - xxmaj wali xxmaj kota xxmaj surabaya xxmaj tri xxmaj rismaharini memilih
merayakan pergantian tahun baru secara sederhana.
xxmaj bersama beberapa pejabat lainnya di xxmaj balai xxmaj kota, mereka memantau kondisi kota melalui lay
xxmaj selama menunggu pergantian tahun, xxmaj risma duduk di bawah sambil menikmati cemilan bakso goren
secangkir jahe hangat.
xxmaj sesekali ia bercerita serta bercanda dengan para pejabat dan wartawan yang hadir.
”xxmaj ayo hitung mundur bersama <number>..<number>..<number>..<number>..<number>,” kata xxmaj r
kepada para beberapa pejabat xxmaj pemkot
di xxmaj balai xxmaj kota, xxmaj minggu <date>.
xxmaj usai menghitung mundur xxmaj risma dan xxmaj wakil xxmaj wali xxmaj kota
xxmaj whisnu xxmaj sakti xxmaj buana dan
beberapa pejabat seperti xxmaj asisten xxmaj I, xxmaj kasatpol xxmaj pp, xxmaj kabag xxmaj umum
dan xxmaj kabag xxmaj humas
langsung bertepuk tangan dan
membaca surat Al Fatiha yang dipimpin Risma.
xxmaj foto: xxmaj zaenal xxmaj Effendi
xxmaj setelah itu, xxmaj wakil xxmaj wali xxmaj kota xxmaj whisnu langsung bersalaman
dan mengucapkan selamat tahun baru pada xxmaj risma
diikuti para pejabat xxmaj pemkot xxmaj surabaya.
xxmaj kembang api yang dinyalakan warga di depan xxmaj balai xxmaj kota xxmaj surabaya
pun bersautan menarik perhatian xxmaj wakil xxmaj wali xxmaj kota xxmaj whisnu xxmaj sakti dan
beberapa pejabat untuk menyaksikan.
xxmaj sedangkan xxmaj risma langsung bergegas meninggalkan xxmaj balai xxmaj kota untuk patroli keliling kc
”xxmaj aku patroli kota jangan ada yang ikuti. xxmaj silahkan bapak ibu pulang,” pamit xxmaj risma pada pej
xxmaj foto: xxmaj zaenal xxmaj effendi xxeos

Table 3. A sample preprocessed news article before tokenization (long sentences have been split in order to fit the column width)

During training and evaluation processes, our BPPT dataset is divided into train set dan validation set. Each train set and validation set have 4 (four) folders to cover four categories elaborated in Table 4. In order to facilitate hyperparameter tuning, the setup of news provider dataset is slightly different from the BPPT dataset. News provider is divided into 3 (three) sets that are train set, validation set, and *test set* as explained in Table 5.

	Train set	Validation set	Test set	Total
News documents	437.583	9.113	8.930	455.626

Table 5. Number of documents broken down into train, validation, and test sets in order to train ULM-FiT

BPPT Document	**ECO**	**INT**	**SCI**	**SPO**
Train set	5.890	5.978	5.720	4.035
Test set	654	664	635	448
Total documents	6.544	6.642	6.355	4.483

Table 4. Detailed number of documents employed in training and testing processes for document classification

Before the ULMFiT model is constructed, the dataset needs preprocessing. A tokenizer from `fast.ai` library, specifically `WordTokenizer` class is employed and produces several special tokens which indicate particular positions of tokens such as xxbos (the beginning of a document), xxmaj (the next word begins with an uppercase letter), xxeos (the end of a document) [19]. Additionally, dates and numbers are converted into <date> and <number> respectively. Fig. 1 shows a sample of

preprocessing results. After preprocessing is done, the results are tokenized into tokens separated by a whitespace (whitespaces); therefore, each document is represented by a list of tokens.

3.2 Steps of building ULM-FiT model

After lists of tokens are obtained, numericalization process is carried out. The process of mapping from tokens into integers is named numericalization process. Basically, the process constructs a list of all possible tokens (a vocab) and replaces each token with its index in the vocab. Having all documents represented as lists of integers, each list of integers is divided into 64 batches (batch size=64). These batches are now ready to be converted into embeddings by using pre-trained language model (LM). Specifically, the pre-trained LM [30] is trained from news articles and illustrated by the first LM in Fig. 5. Next, the embeddings are fed into a recurrent neural network (RNN), using a AWD-LSTM architecture.

The next step is fine-tuning the AWD-LSTM architecture on BPPT dataset. This process is needed because there is a possibility that the distribution of BPPT dataset is different from news provider dataset. We freeze the architecture and train only the embeddings. After training the embeddings, we unfreeze the architecture and train both embeddings and the architecture. Furthermore, we save all of the model except the final layer; the final layer contains an activation layer for computing the probabilities of picking each token in the vocab. We name the model except the final layer as encoder [19]. The encoder itself is represented as the second LM in Fig. 5.

The final step is fine-tuning a text classifier on BPPT dataset again; however, this time we let the classifier learn document categories of the dataset. Specifically, we load the encoder and train the classifier with discriminative fine-tuning and gradual unfreezing [21]. Moreover, we unfreeze several layers as the gradual unfreezing in NLP makes a significant difference when a few layers are frozen instead of all layers which is a common best practice in computer vision. Finally, we finish building our ULM-FiT model; the classifier in Fig. 5 illustrates the text classifier.

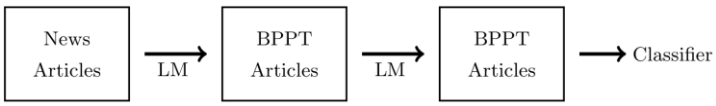


Fig. 5 Three steps of constructing ULMFiT model (LM=Language Model)

The Hyperparameters	
batch size	64
Batch size of BPPT	70
Embedding size	300
Number of layers in embedding	2
Number of hidden activations per layer	300
Initial learning rate	5×10^{-3}

Table 6 . The recommended hyperparameters for AWD-LSTM

The hyperparameters for AWD-LSTM model in our experiments are recommended from the original paper [21] as depicted in Table [6].

3.3 Baseline models

Baseline models are needed as to examine the effect of ULM-FiT to reduce overfitting problem. The baseline models constructed as a comparison are neural networks with two hidden layer; specifically, each layer has eight nodes. The number is chosen because more than that number shall make the neural nets overfitting on the dataset. Moreover, each node in hidden layers has relu activation function. The decision to take this activation function is that this activation has capability to mitigate the vanishing gradient problem [34]. With equally the same reason, we also specify number of inputs are the most frequent 10.000 words and number of outputs are four categories, that are economics, international, science, and sports. Number of instances in validation set is 1.000 and the rest is considered as test set.

4. Research Methodology

Fig. 6 shows the training and validation error from the baseline model during 20 epochs. The training error reaches an almost perfect score, 99% whereas the validation error is 95%. The difference between train and validation set is 4%, which is quite large. This signifies an overfitting in the model. We shall see how ULMFiT model overcomes this major problem.

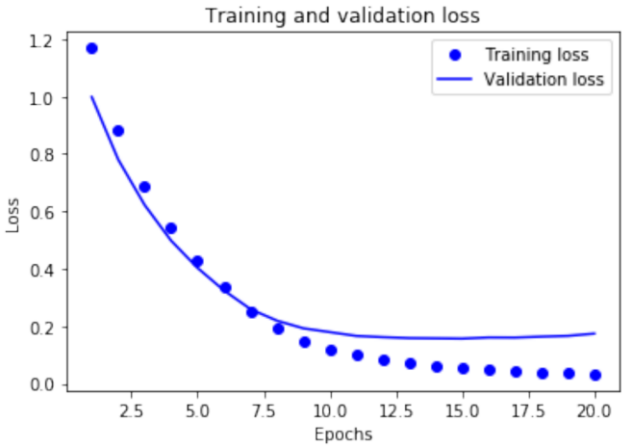


Fig. 6 . Training and validation error from the baseline model

4.1 Validation performance of ULMFiT

Table 7 shows the training process of constructing the pre-trained LM based on pointer sentinel mixture architecture [30] on news articles dataset. This architecture acts as a pre-trained LM (the first LM) as described in Fig. 5. The learning runs smoothly as training and validation losses are getting smaller and smaller. Each epoch approximately takes 2.5 hours.

Since both vocabularies of news provider and BPPT dataset are different, there might be several words in BPPT dataset which are not in the pre-trained model vocabulary. Consequently, the embeddings of words in the pre-trained model are merged with random embeddings for words which are not in the pre-trained model

vocabulary. In order to merge both embeddings, we fine-tune the model by freezing the pre-trained model but not the embeddings. Table 8 shows the training and validation loss of this fine-tuning.

epoch	train loss	valid loss	time
0	5.116353	4.616845	2:35:35
1	4.353452	4.481172	2:36:02
2	4.277426	4.473105	2:36:07
3	4.234690	4.434146	2:36:19
4	4.186194	4.386984	2:36:17
5	4.129720	4.330991	2:36:18
6	4.066865	4.276275	2:36:24
7	4.006233	4.224578	2:36:07
8	3.958543	4.196958	2:33:50
9	3.931666	4.187688	2:33:24

Table 7 . Results of training the LM on news provider dataset; specifically, this LM acts as the pre-trained LM

epoch	train loss	valid loss	time
0	5.914727	4.485229	00:11

Table 8 . Results of training while the pre-trained model is frozen except the embeddings

Next, the merged embeddings are fed into the state-of-the-art language model AWD-LSTM [28] and trained with discriminative learning rates whose initial values equal to 2×10^{-3}. Importantly, discriminative learning rates are utilized because different layers bring in different kinds of information [46,21]. Therefore, each layer has a different learning rate. The discriminative learning are combined with Stochastic Gradient Descent (SGD) technique combined and are stated as follows:

$$\Theta_t^{[l]} = \Theta_{t-1}^{[l]} - \alpha^{[l]} \cdot \nabla_{\Theta^{[l]}} J(\Theta) \qquad (9)$$

with t is a time step, [l] is lth layer, and $\alpha^{[l]}$ is a learning rate for the lth layer. The results of training with discriminative learning rates are shown in Table 9. Finishing the training, we have an updated LM shown as the second LM in Fig. 5.

After training AWD-LSTM model, we save the LM model except the final layer named as an encoder. Next, this encoder is loaded before fine-tuning a classifier which becomes the final stage of ULM-FiT construction. The fine-tuning process adds two linear blocks on the model. Each block employs batch normalization [22] and dropout with ReLU activation functions for intermediate layers and a softmax activation for showing the probability distributions of every class in the last layer. Critically, input data for the final layer is the output of the last hidden layer, the average of all outputs from all hidden states, and the maximum of all outputs from hidden states. These three inputs are merged by a technique so-called concat pooling.

epoch	train loss	valid loss	time
0	4.561834	4.385835	00:12
1	4.446705	4.267622	00:12
2	4.307850	4.146501	00:12
3	4.176742	4.046482	00:12
4	4.070283	3.973104	00:12
5	3.990463	3.918991	00:12
6	3.929258	3.884007	00:12
7	3.876218	3.862094	00:12
8	3.853766	3.854638	00:12
9	3.837198	3.852936	00:12

Table 9 . The training and validation loss during training AWD-LSTM with discriminative learning rates

We train the model with discriminative learning rates and gradual unfreezing. Firstly we unfreeze the last two parameter groups; secondly, we unfreeze the last three parameter groups; lastly, we unfreeze the whole model. This NLP best practice is slightly different from the one of computer vision which considers unfreezing the complete model at once. However, this NLP best practice proves significantly making improvements [19]. Furthermore, we opt for choosing a small number of epochs conforming to the original paper's setting [19] due to the so-called super-convergence phenomenon [40]. Table 10, 11, and 12 display the three gradual unfreezing processes respectively.

epoch	train loss	train acc	valid loss	valid acc	time
0	0.360814	0.876289	0.232974	0.920033	00:05

Table 12 . Results of training with gradual unfreezing the last two parameter groups

epoch	train loss	train acc	valid loss	valid acc	time
0	0.275453	0.904454	0.184063	0.930862	00:06

Table 11 . Results of training with gradual unfreezing the last three parameter groups

epoch	train loss	train acc	valid loss	valid acc	time
0	0.211038	0.927253	0.175658	0.934194	00:08
1	0.182362	**0.936318**	0.169076	**0.933778**	00:09

Table 10 . Results of training with gradual unfreezing the whole model

Specifically, Table 12 shows the prediction accuracy of ULMFiT. Utilizing ULMFiT produces a difference value between train and valid accuracies which approaches zero (2.5×10^{-3}). This remarkable result suggests ULMFiT should be the excellent choice of model to overcome overfitting.

5. Conclusion

This research explores the use of transfer learning technique to leverage language model in order to do a text classification task. Particularly, we implement Universal Language Model Fine-tuning (ULMFiT) for Indonesian language as this technique has the capability to overcome overfitting problem. To the best of our knowledge, we believe this is the first ULMFiT model fine-tuned to Indonesian news articles. Although the prediction accuracy in validation set is lower than the baseline method, ULMFiT successfully reduces the overfitting difference between training and validation accuracies from 4% to nearly zero.

References

1. Altınel, B., Ganiz, M.C., Diri, B.: A corpus-based semantic kernel for text classification by using meaning values of terms. Engineering Applications of Artificial Intelligence 43, 54-66 (2015). https://doi.org/10.1016/j.engappai.2015.03.015,http://www.sciencedirect.com/science/article/pii/S0952197615000809
2. Anshori, M., Mahmudy, W.F., Supianto, A.A.: Classification tuberculosis dna using lda-svm. Journal of Information Technology and Computer Science 4(3), 233–240 (2019). https://doi.org/10.25126/jitecs.201943113
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv e-prints abs/1409.0473 (Sep 2014), https://arxiv.org/abs/1409.0473
4. Baldi, P., Sadowski, P., Whiteson, D.: Searching for exotic particles in high-energy physics with deep learning. Nature communications 5, 4308 (2014)
5. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2), 157–166 (1994)
6. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800-1807 (2017)
7. Chollet, F.: Deep Learning with Python 1st Edition. Manning Publications (2017)
8. CireşAn, D., Meier, U., Masci, J., Schmidhuber, J.: Multi-column deep neural network for traffic sign classification. Neural networks 32, 333–338 (2012)
9. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. arXiv preprint arXiv:1301.3572 (2013)
10. Dahl, G.E., Jaitly, N., Salakhutdinov, R.: Multi-task neural networks for QSAR predictions. CoRR abs/1406.1231 (2014), http://arxiv.org/abs/1406.1231
11. Eisenstein, J.: Introduction to Natural Language Processing. Adaptive Computation and Machine Learning series, MIT Press (2019), https://books.google.co.id/books?id=72yuDwAAQBAJ
12. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence 35(8), 1915–1929 (2013)
13. Finn, C., Tan, X.Y., Duan, Y., Darrell, T., Levine, S., Abbeel, P.: Learning visual feature spaces for robotic manipulation with deep spatial autoencoders. CoRR abs/1509.06113 (2015), http://arxiv.org/abs/1509.06113
14. Gershgorn, D.: How to develop lstm models for time series forecasting. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/ (2017), accessed: 2019-01-15
15. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
17. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech

recognition: The shared views of four research groups. IEEE Signal processing magazine 29(6), 82–97 (2012)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
19. Howard, J., Gugger, S.: Deep Learning for Coders with fastai & PyTorch AI Applications without a PhD. O'Reilly (2020)
20. Howard, J., Ruder, S.: Introducing state of the art text classification with universal language models. http://nlp.fast.ai/classification/2018/05/15/introducting-ulmfit.html (2018), accessed: 2019-01-15
21. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 328–339 (2018)
22. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Bach, F., Blei, D. (eds.) Proceedings of the 32nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 37, pp. 448–456. PMLR, Lille, France (07–09 Jul 2015), http://proceedings.mlr.press/v37/ioffe15.html
23. Knowles-Barley, S., Jones, T.R., Morgan, J., Lee, D., Kasthuri, N., Lichtman, J.W., Pfister, H.: Deep learning for the connectome. In: GPU Technology Conference. Vol. 26 (2014)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
25. Lewis-Kraus, G.: The great a.i. awakening. https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html (2016), accessed: 2019-01-15
26. Liu, B.: Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications). Springer (2013)
27. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to information retrieval (2008)
28. Merity, S., Keskar, N.S., Socher, R.: Regularizing and optimizing LSTM language models. CoRR abs/1708.02182 (2017), http://arxiv.org/abs/1708.02182
29. Merity, S., McCann, B., Socher, R.: Revisiting activation regularization for language rnns. CoRR abs/1708.01009 (2017), http://arxiv.org/abs/1708.01009
30. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. CoRR abs/1609.07843 (2016), http://arxiv.org/abs/1609.07843
31. Mirończuk, M.M., Protasiewicz, J.: A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications 106, 36 - 54 (2018). https://doi.org/10.1016/j.eswa.2018.03.058, http://www.sciencedirect.com/science/article/pii/S095741741830215X
32. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature 518(7540), 529 (2015)
33. Ng, A.Y.: Machine Learning Yearning: Technical Strategy for AI Engineers, In the Era of Deep Learning. deeplearning.ai (2018)
34. Nielsen, M.A.: Neural networks and deep learning, vol. 25. Determination press San Francisco, CA, USA: (2015)
35. Oktaria, A.S., Prakasa, E., Suhartono, E.: Wood species identification using convolutional neural network (cnn) architectures on macroscopic images. Journal of Information Technology and Computer Science 4(3), 274–283 (2019). https://doi.org/10.25126/jitecs.201943155
36. Pinheiro, R.H., Cavalcanti, G.D., Tsang, I.R.: Combining dissimilarity spaces for text categorization. Information Sciences 406-407, 87 – 101 (2017). https://doi.org/10.1016/j.ins.2017.04.025, http://www.sciencedirect.com/science/article/pii/S0020025517306722
37. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.: Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. CoRR abs/1801.04381 (2018), http://arxiv.org/abs/1801.04381

38. Sermanet, P., Kavukcuoglu, K., Chintala, S., LeCun, Y.: Pedestrian detection with unsupervised multi-stage feature learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3626–3633 (2013)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1409.1556
40. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. CoRR abs/1708.07120 (2017), http://arxiv.org/abs/1708.07120
41. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in neural information processing systems. pp. 3104-3112 (2014)
42. Szegedy, C., Wei Liu, Yangqing Jia, Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9 (2015)
43. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A Survey on Deep Transfer Learning: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III, pp. 270–279 (10 2018)
44. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R.: Regularization of neural networks using dropconnect. In: International Conference on Machine Learning. pp. 1058–1066 (2013)
45. Wang, D., Wu, J., Zhang, H., Xu, K., Lin, M.: Towards enhancing centroid classifier for text classification: a border-instance approach. Neurocomputing 101, 299-308 (2013). https://doi.org/10.1016/j.neucom.2012.08.019, http://www.sciencedirect.com/science/article/pii/S0925231212006595
46. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 3320–3328. Curran Associates, Inc. (2014), http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf