



ONE WEEK FACULTY DEVELOPMENT PROGRAMME ON
Artificial Intelligence and Machine Learning
(2nd August to 6th August 2022)

Natural Language Processing and Application

Dr. Shantipriya Parida

Agenda

- Overview
- Deep Learning in NLP
- Case Studies
 - Case Study 1 - Language Corpus Development
 - Case Study 2 - Machine Translation
 - Case Study 3 - Text Summarization
 - Case Study 4 - Language Detection
 - Case Study 5 - Fake News Detection
 - Case Study 6 - Operant Motive Classification
 - Case Study 7 - Language Model Development
- Conclusion

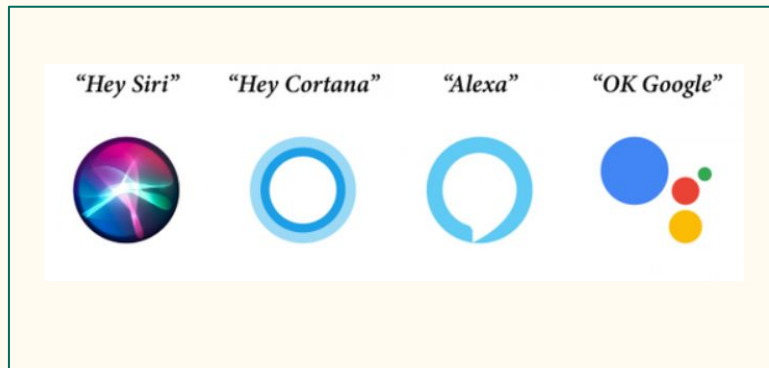
Overview

- Natural language processing (NLP) helps computers communicate with humans in their own language and scales other language-related tasks.
- NLP makes it possible for computers to read text, hear speech, interpret it, measure sentiment and determine which parts are important.



NLP in Daily Life

- Personal assistants: Siri, Cortana, and Google Assistant.
- Auto-complete: In search engines (e.g. Google).
- Spell checking: Almost everywhere, in your browser, your IDE (e.g. Visual Studio), desktop apps (e.g. Microsoft Word).
- Machine Translation: Google Translate.
- Chat bots.



Why NLP is difficult ?

- **Natural language** is highly ambiguous.
- Words can have several meanings and contextual information is necessary to correctly interpret sentences.
- Syntactic analysis (syntax) and semantic analysis (semantic) are the two primary techniques that lead to the understanding of natural language.



(a) Street sign advising of penalty.

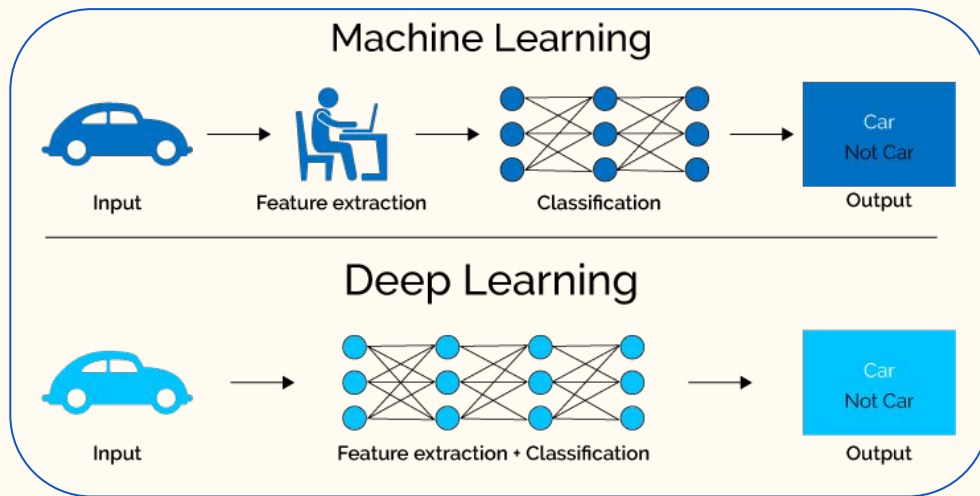


(b) The penalty box is white lined.

Figure: An illustration of two meanings of the word “penalty” exemplified with two images

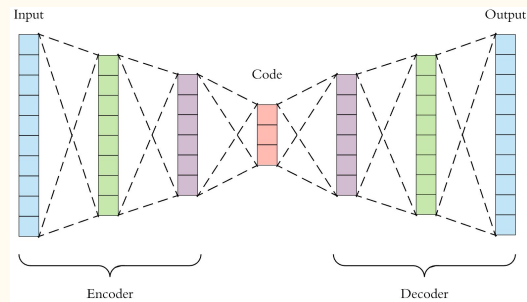
Deep Learning in NLP

- A machine learning subfield of learning **representations** of data.
- Exceptionally effective at **learning patterns**.
- Deep learning algorithms attempt to learn (multiple levels of) representation by using a **hierarchy of multiple layers**.
- If you provide the system **tons of information**, it learns to respond in useful ways.



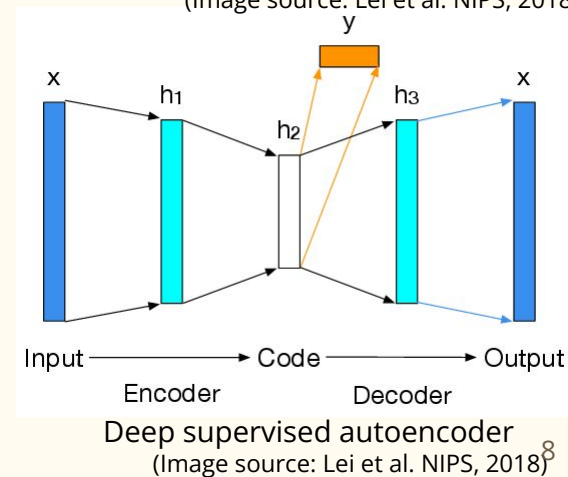
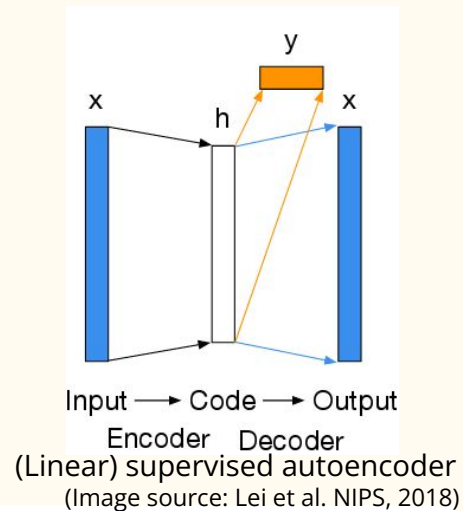
Deep Learning for NLP: Autoencoders

- Learning good representations lies at the core of Deep Learning (DL)
- Over the last few years, DL has made amazing advances in NLP
- Recently, autoencoders represent an alternative to contrastive unsupervised word learning
 - Are able to learn both linear and non-linear transformations
- Autoencoders can discover low-dimensional, less sparse, and robust features for classification



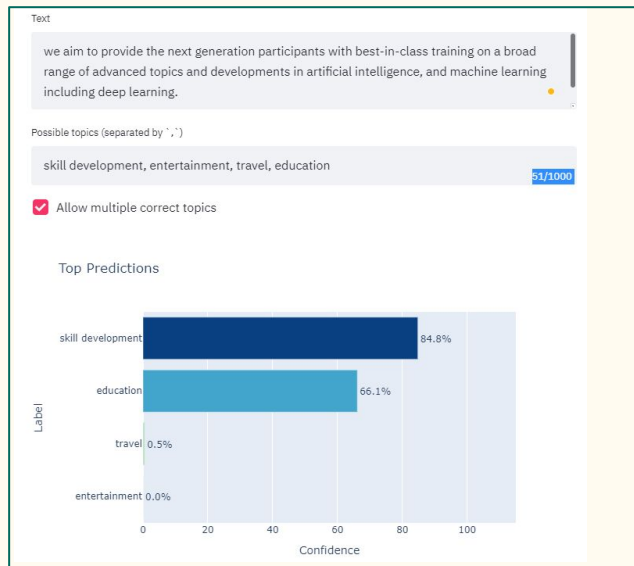
Supervised AutoEncoder

- A **supervised autoencoder (SAE)** is an autoencoder with the addition of a supervised loss on the representation layer
- The addition of supervised loss to the autoencoder acts as a regularizer and results better representation for the desired task
- Although **SAE** have been tested on many image classification tasks, they have **not** been extensively tested on **NLP** tasks



Text Classification

- Text classification is a task of NLP where the model needs to predict the classes of the text documents.
- In the traditional process, we are required to use a huge amount of labelled data to train the model, and also they can't predict using the unseen data.
- Adding zero-shot learning with text classification has taken NLP to the extreme. <https://huggingface.co/zero-shot/>
- Zero-shot text classification technique classify the text documents without using any single labelled data or without having seen any labelled text.



Zero-shot Text Classification

Named Entity Recognition

- **Named entity recognition** is a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.
- Entities can be: Organizations, Quantities, Monetary values, Percentages, People's names, Company names, Geographic locations (Both physical and political), Product names, Dates and times, Amounts of money, Names of events.

Original Text

The program is planned for the summer holiday in Odisha (June-July 2022) and will be in virtual mode considering teaching by many international experts. The participants will be based on registration considering the eligibility and background of the participants. The maximum number of participants will be 100. The AI ML Summer School is a five days program and will be conducted on weekends (Saturday and Sunday) only with 3 lecturer sessions per day and each session will be 1.5 hours (1-hour Theory, 30 minutes Demo) except first and last sessions. However, students can execute the assigned mini-project on other days at their convenient time in groups.

Analysis Result

the summer holiday/DATE

Odisha/GPE

June-July 2022/DATE

100/CARDINAL

The AI ML Summer School/ORG

five days/DATE

Saturday/DATE

Sunday/DATE

3/CARDINAL

1.5 hours/TIME

1-hour/TIME

30 minutes/TIME

Topic Modelling

- What is Topic Modelling ?
 - Topic modeling is a statistical modeling approach to discover the abstract “topics” occurs in a collection of documents.
- Types of Topic Modeling
 - Unsupervised, and Semi-supervised
- Application of Topic Modeling
 - text mining, text classification, machine learning, information retrieval, and recommendation engines.

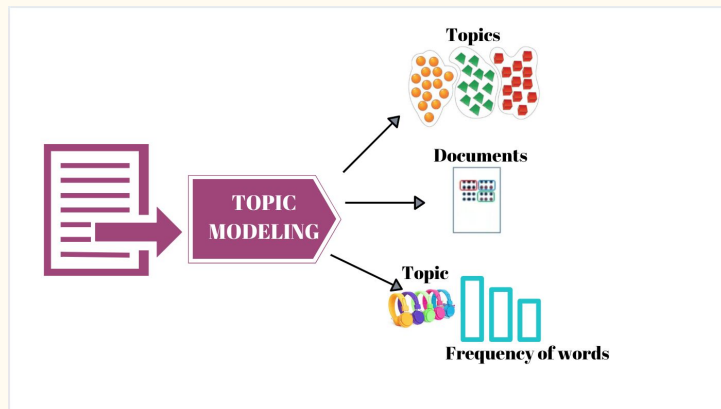
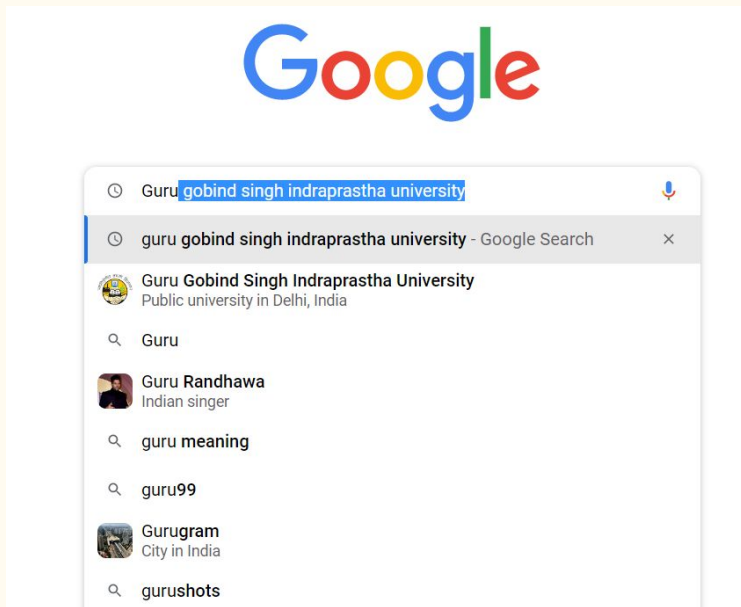


Fig: Topic Modeling

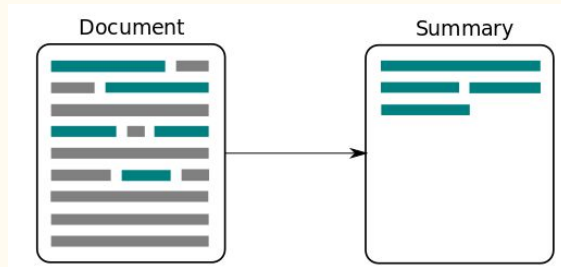
Language Model

- Language modeling is the task of predicting the next word in a sentence, given all previous words.
- It is an effective task for using unlabeled data to pretrain neural networks in NLP.
- Language models capture general aspects of the input text that is almost universally useful.



Text Summarization

- Automatic text summarization aims to transform lengthy documents into shortened versions, something which could be difficult and costly to undertake if done manually.
- Two major approaches for automatic summarization are: extractive and abstractive.
- The extractive summarization approach produces summaries by choosing a subset of sentences in the original text.
- The abstract text summarization approach aims to shorten the long text into a human readable form that contains the most important fact from the original text

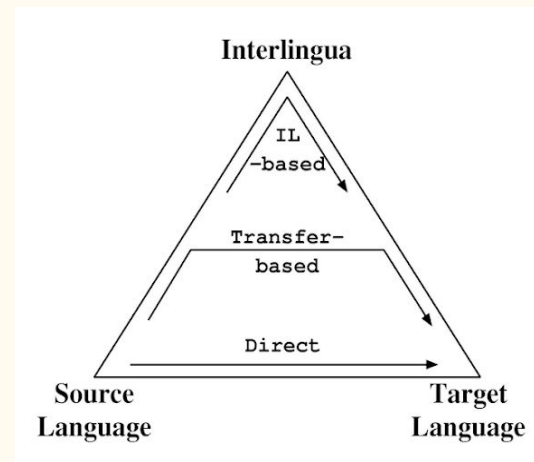


Machine Translation

- Automatic conversion of text/speech from one natural language to another.

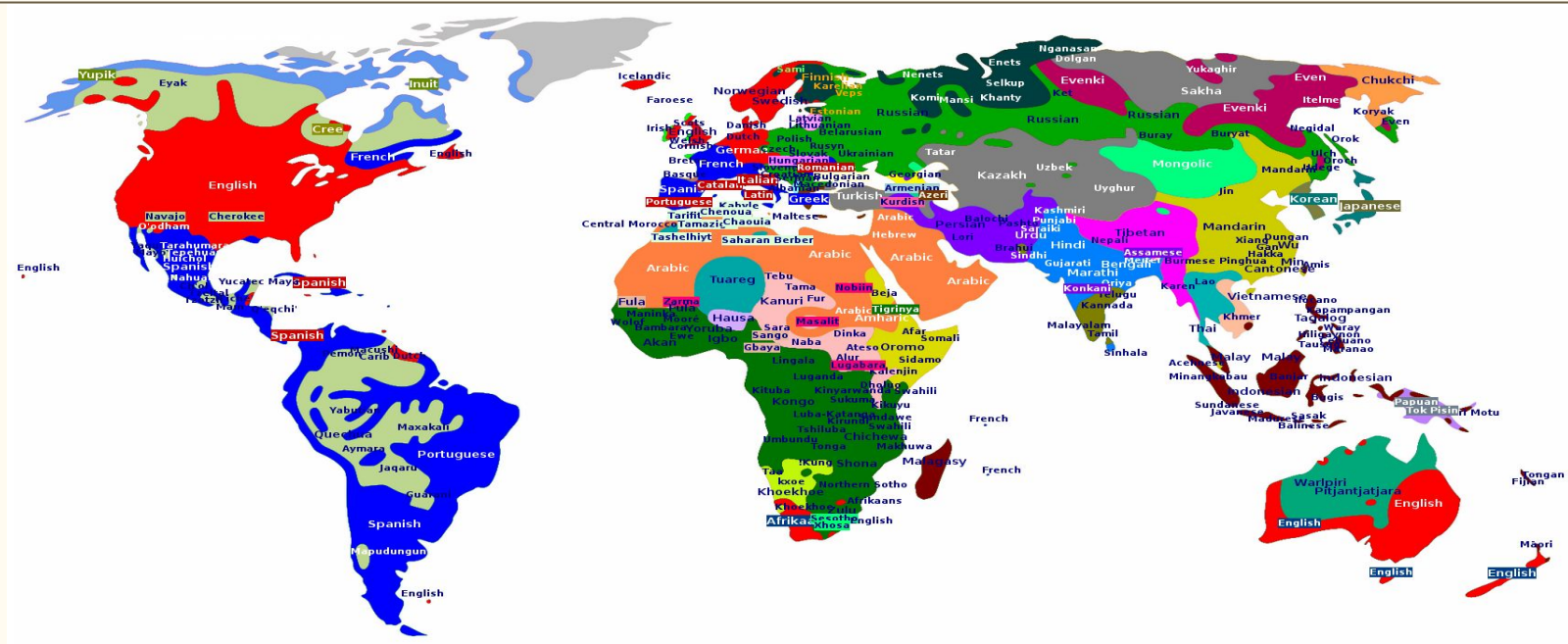


- Machine translation approaches:
 - Grammar-based
 - Interlingua-based
 - Transfer-based
- Direct
 - Example-based
 - Statistical
 - Neural



Case 1 - Corpus Development

- The [ethnologue.com](https://en.wikipedia.org/wiki/Template:Distribution_of_languages_in_the_world) website lists over 7000 languages in the world.



The geographical pattern of the major languages of the world.
Source: https://en.wikipedia.org/wiki/Template:Distribution_of_languages_in_the_world

Case 1 - Corpus Development

Off the Coast of India, Another Language Dies

By Ishaan Tharoor | Wednesday, Feb. 17, 2010

Нравится 243 Tweet

Share

Read Later

On some days, Boa Sr would sit silently in the jungle surrounding her home on one of India's Andaman Islands and gaze up at the sky. According to researchers who looked on, birds perched above would descend to the ground and inspect her; in turn Boa Sr spoke to them in her native tongue, calling them her ancestors and her friends. Her speech was rich with words of the natural world, words of the forest and the sea that some linguists suspect date back tens of thousands of years to the first migrations of man. Boa Sr was the last person alive to know them. In early February, she passed away, leaving behind no surviving siblings or children. As she died, so too did the language of her people.



Alok Das/ SURVIVAL INTERNATIONAL / AFP

Boa Sr, the last speaker of Bo, one of the 10 Great Andamanese languages, on the Andaman and Nicobar Islands

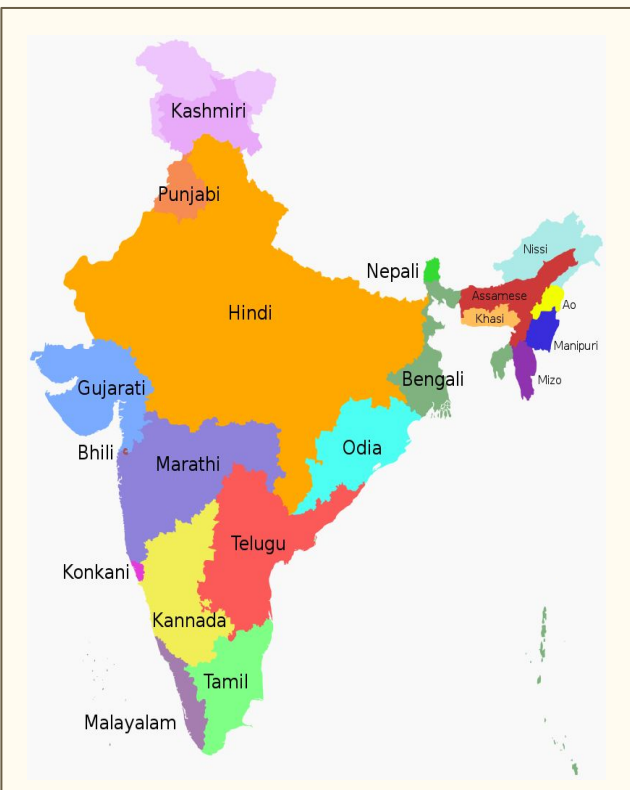


Image source:
https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India#/media/File:Language_region_maps_of_India.svg

source: <http://content.time.com/time/world/article/0,8599,1964610,00.html>

Need for Language Resource

- Wikipedia has texts in 313 languages.
- Natural language technology development depends on large numbers of language resources (text / speech).
- Lack of language resources affects the development of natural language technologies.



Corpus

- **Corpus (plural corpora)** : A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech.
- A **corpus** can be made up of everything from newspapers, novels, recipes and radio broadcasts to television shows, movies and tweets.
- In NLP, a corpus contains text and speech data that can be used to train AI and machine learning systems.
- Generally, the larger the size of a corpus, the better (prioritize quantity over quality).



Corpus - How to Build ?

- Data Collection
 - Data type
 - Text/Image/Speech/Video
 - Identify source
 - Web, Social Media, Books, Recordings
 - Web scraping
 - Identify URLs (e.g. language, text, tags)
 - Bots
 - Optical Character Recognition (OCR)
 - Extract data
 - tools: Python, BeautifulSoup
- Data Processing
 - segmentation, alignment
 - Purnaviram, Hunalign
- Finalization and Release
 - Split train/dev/test set
 - Baseline
 - License
 - Release platform
 - Share/organize shared task
 - WMT, WAT, ICON, etc...



Image source:
<https://medium.com/analytics-vidhya/web-scraping-and-coursera-8db6af45d83f>

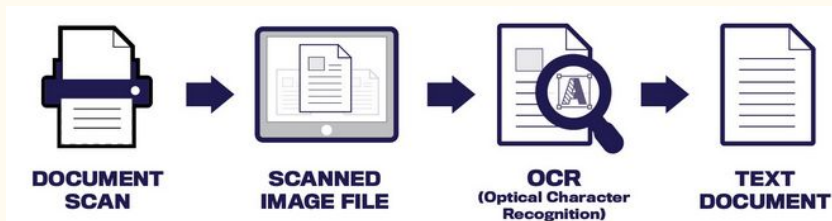
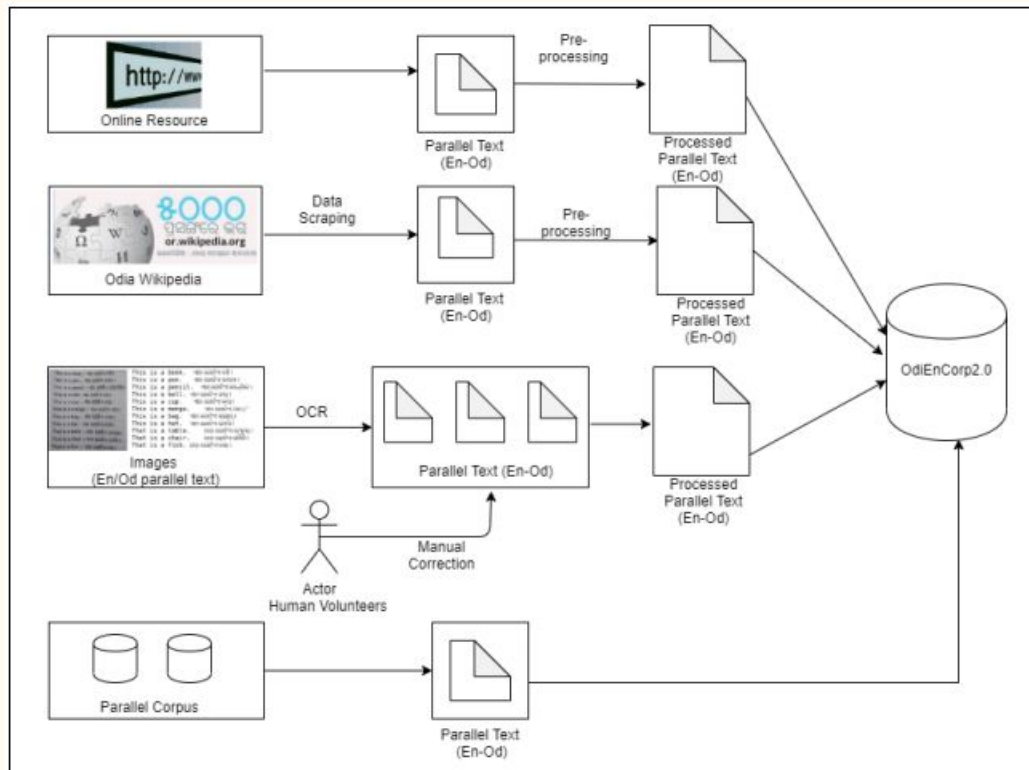


Image source: Image source:
<https://medium.com/states-title/using-nlp-bert-to-improve-ocr-accuracy-385c98ae174c>

Sample (OdiEnCorp)

- Data extracted from other online resources.
- Data extracted from Odia Wikipedia.
- Data extracted using Optical Character Recognition (OCR).
- Data reused from existing corpora.



Sample (OdiEnCorp)

- Data Processing

- Extraction of plain text.
 - Python script to scrape plain text from HTML page.
- Manual processing.
 - Correction of noisy text extracted using OCR-based approach.
- Sentence segmentation.
 - Paragraph segmented into sentences based on English full stop (.) and Odia Danda (।) or Purnaviram.
- Sentence alignment.
 - Manual sentence alignment for Odia Wikipedia articles where text in two language are independent of each other.

Dataset	#Sentences	#Tokens	
		EN	OD
Train 2.0	69260	1340371	1164636
Dev 2.0	13429	157951	140384
Test 2.0	14163	185957	164532

Dataset Statistics.

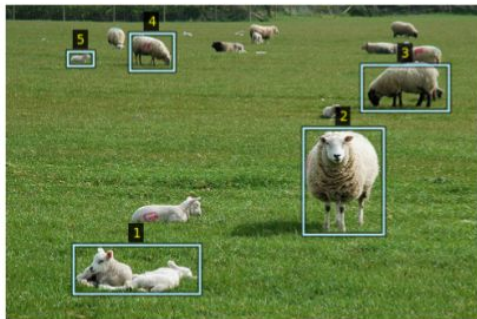
Availability

OdiEnCorp 2.0 is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, CC-BY-NC-SA at :

<http://hdl.handle.net/11234/1-3211>

Motivation

Do Visual Context Disambiguates ?



Caption 1: Two lambs lying in the sun.

Hindi MT: दो भेड़ के बच्चे सूरज में झूठ बोल रहे हैं

Gloss: Two baby sheep are **telling lies** ...

Selected surrounding captions:

2. Sheep standing in the grass
3. Sheep with black face and legs
4. Sheep eating grass
5. Lamb sitting in grass.

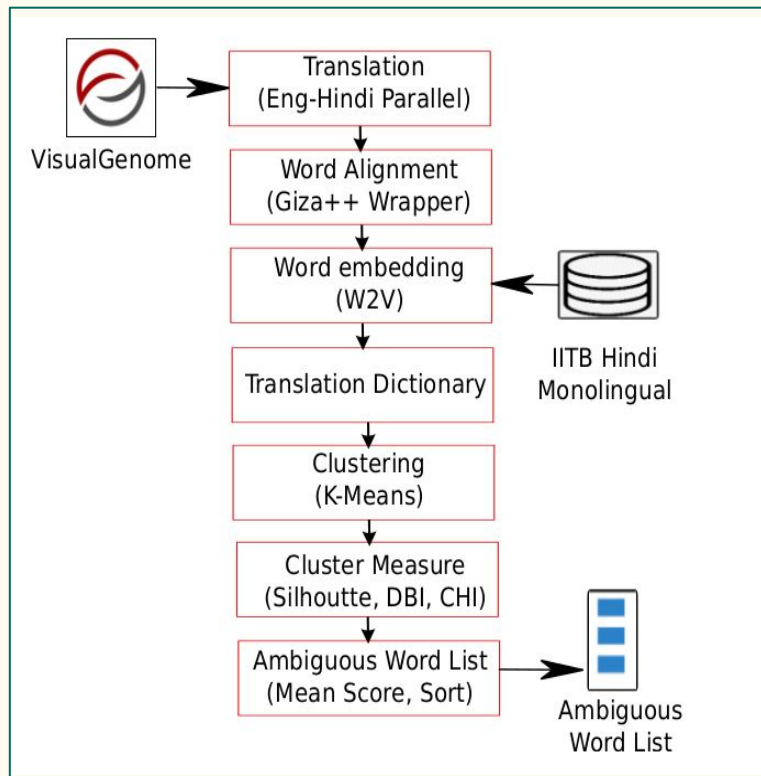
Multimodal Corpus

- Multi-modal content is gaining popularity in machine translation (MT) community due to its appealing chances to improve translation quality.
- It has application in commercial application
 - Translation of image captions in online news article
 - Machine translation of e-commerce product listings.
- Although neural machine translation (NMT) models very good for large parallel texts, some inputs can remain genuinely ambiguous, especially if the input context is limited.
 - Exa: “**mouse**” in English (source) which can be translated into different words in Hindi based on the context (e.g. either a computer mouse or a small rodent)

Steps (Training and Test)

- The starting point were 31,525 randomly selected images from Visual Genome
- We translated all 31,525 captions into Hindi using the NMT model (Tensor-to-Tensor)
- We uploaded the image, the source English caption and its Hindi machine translation into a “Translation Validation Website”
- Volunteers post-edited all the Hindi translations.
- We manually verified and finalized the post-edited files to obtain the training and test data.

Steps (Challenge Test set)







Overall pipeline for ambiguous word finding from input corpus

1. Translate all English captions of visual Genome (3.15 million unique strings) using Google translate.
2. Apply word alignment on the synthetic parallel corpus using GIZA++ Wrapper.
3. Extract all pairs of aligned words in the form of a “translation dictionary”. Dictionary contains key/value pairs of the English word (E) and all its Hindi translations ($H_1, H_2, \dots H_n$), $E \rightarrow \{H_1, H_2, \dots H_n\}$.
4. Train Hindi word2vec (W2V) word embeddings. We used the gensim implementation and trained it on IITB Hindi Monolingual Corpus which contains about 45 million Hindi sentences.
5. For each English word from the translation dictionary, get all Hindi translation words and their embeddings.
6. Apply K-means clustering algorithm to the embedded Hindi words to organize them according to their word similarity.
7. Evaluate the obtained clusters with the Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harbaz Index (CHI).
8. Sort the list in descending order to get the most ambiguous words.
9. Manually check the list and extract the most ambiguous English words.

Challenge Test set

	Word	Segment Count
1	Stand	180
2	Court	179
3	Players	137
4	Cross	137
5	Second	117
6	Block	116
7	Fast	73
8	Date	56
9	Characters	70
10	Stamp	60
11	English	42
12	Fair	41
13	Fine	45
14	Press	35
15	Forms	44
16	Springs	30
17	Models	25
18	Forces	9
19	Penalty	4
	Total	1400

Challenge test set: ambiguous words

	<p>English Input: gold religious cross on top of golden ball</p> <p>Translated Output: सोने की गेंद के शीर्ष पर स्वर्ण धार्मिक क्रॉस .</p> <p>Gloss: Gold religious cross on top of golden ball</p>
	<p>English Input: a blue wall beside tennis court</p> <p>Translated Output: टेनिस कोर्ट के पास एक नीली दीवार हैं ।</p> <p>Gloss: Blue wall near the tennis court</p>
	<p>English Input: the tennis court is made up of sand and dirt</p> <p>Translated Output: टेनिस कोर्ट रेत और गंदगी से बनी है।</p> <p>Gloss: Tennis court is made of sand and dirt</p>
	<p>English Input: A crack on the court</p> <p>Translated Output: अदालत पर एक crack</p> <p>Gloss: A crack on the <u>judicial court</u></p>

Availability

Hindi Visual Genome



Hindi Visual Genome

Hindi-English Multimodal Dataset

<https://ufal.mff.cuni.cz/hindi-visual-genome>

Hindi Visual Genome 1.0

Used in WAT 2019

Hindi Visual Genome 1.1

Used in WAT
2020,2021,2022

Case 2 - Machine Translation

Multimodal Machine Translation

- **Multimodal Translation** refers to the extraction of information from more than one modality where it is assumed that alternative views would be used for input data.
- There is English-Bengali text-only parallel corpora available for developing machine translation (MT) systems; however, there is no such multimodal dataset for Bengali till now.
- We have provided a Bengali Visual Genome (BVG) dataset in this work that can facilitate research and development of corresponding multimodal as well as image captioning tasks.

BVG Dataset

Dataset	#Sentences	#Tokens	
Train	28930	143156	113993
D-Test	998	4922	3936
E-Test	1595	7853	6408
C-Test	1400	8186	6657

Table1: Number of Tokens for English (EN) and Bengali (BN) for each set are reported; for test dataset, development test (D-Test), evaluation test (E-Test), and challenge test (CTest) sets are prepared similarly as training dataset



Fig2: Sample item from the BVG dataset
English Text: A girl playing tennis.
Bengali Text: একটি মেয়ে টেনিস খেলছে

Description of the MMT System

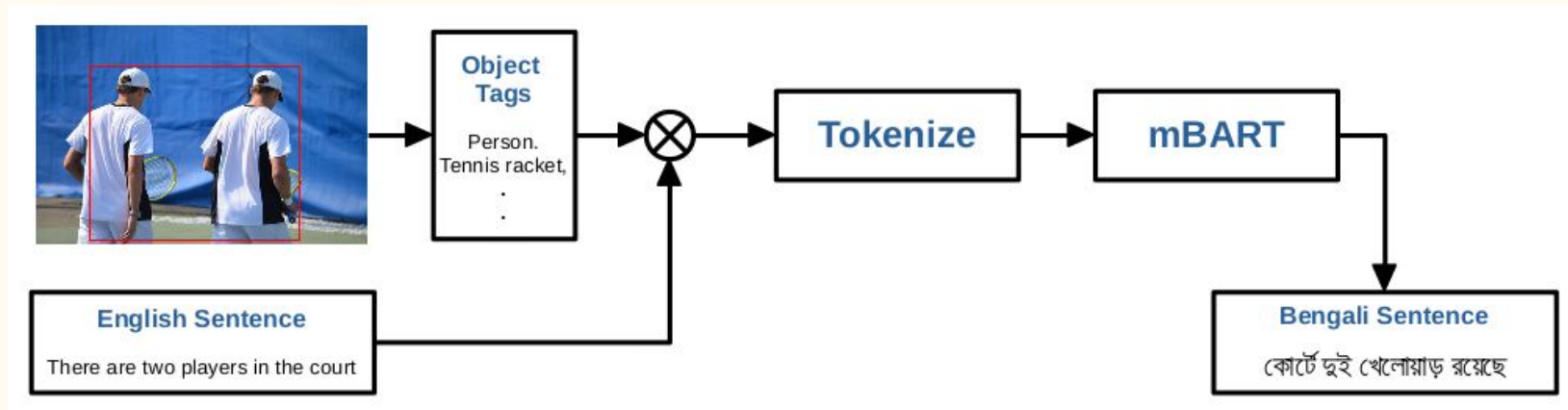


Fig. Multimodal machine translation. The object tags extracted from images along with the English source text input to the mBART to generate the Bengali translation output.

Result

MT System	D-Test BLEU	E-Test BLEU	C-Test BLEU
Text-to-text transformer	42.8	35.6	17.2
Text-to-text mBART	49.8	39.6	25.9
Multimodal mBART	51.1	43.5	26.8

Table: Text only and multimodal translation performance on the BVG dataset

Sample Output




Input Image	Input Caption	Text-to-text Result	MMT Result
	The water bottle on the stand	স্ট্যান্ডে জলের বো- তল	স্ট্যান্ডে জলের বো- তল
		“Water bottle on the stand”	“Water bottle on the stand”
	Two people wait- ing to cross	দুজন লোক ক্রস অপেক্ষা করছে	দুজন লোক ক্রস অপেক্ষা করছে
		“Two people are waiting cross”	“Two people are waiting cross”
	Man standing on a tennis court	টেনিস কোর্টে দাঁড়ি- য়ে লোক	টেনিস কোর্টে দাঁড়ি- য়ে লোক
		“Man standing on a tennis court”	“Man standing on a tennis court”

Fig: Samples of Text-to-text and Multimodal Translation obtained from the Text-to-text mBART and the Multimodal mBART systems

Sample Output



	stamp on boy's left hand	ছেলেটির বাম হাতে স্ট্যাক্স	ছেলেটির বাম হাতে স্ট্যাম্প
		"Stank on boy's left hand" (incorrect Bengali word 'Stank' obtained in T2T translation)	"Stamp on boy's left hand" (correct Bengali word 'stamp' obtained in MMT translation)
	fence around the court	আদালতের চারদিকে বেড়া	কোর্টের চারপাশে বেড়া
		"Fence around the court" (court is translated by T2T as <i>Judicial Court</i> in Bengali)	"Fence around the court" (court is translated by MMT as <i>Tennis Court</i> in Bengali)

Fig: Samples of Text-to-text and Multimodal Translation obtained from the Text-to-text mBART and the Multimodal mBART systems

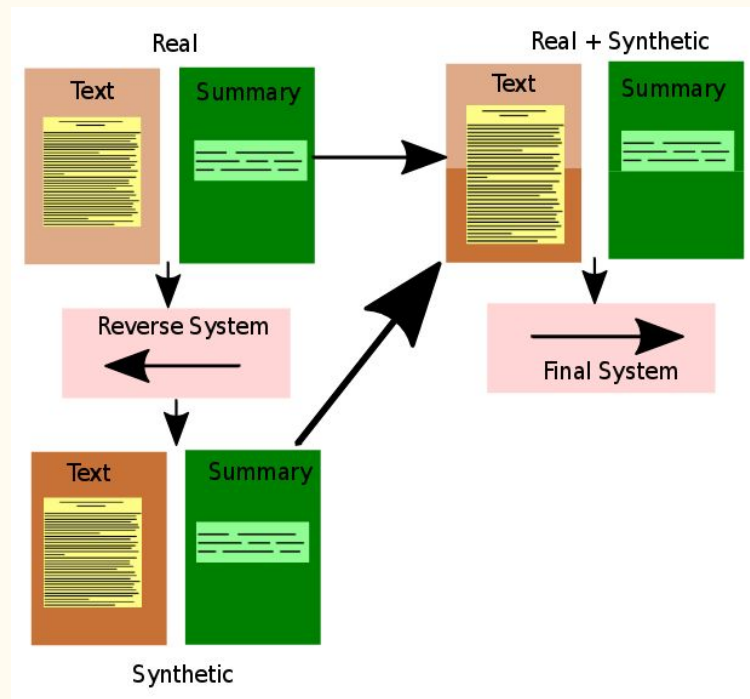
Case 3 - Text Summarization

(Usage of Synthetic Data for Text Summarization)

- Based on Idiap participation in the SwissText 2019 challenge (100'000/2'000) paragraphs and summaries for training/evaluation.
- *Use of synthetic data*: a popular approach in machine translation for the low resource conditions to improve the quality.
- Can such approaches work for the text summarization task ?.

Method

- Use a state-of-the-art “Transformer Model” as implemented in OpenNMT-py.
- Different experiments performed based on real and synthetic data.
- Synthetic data used to increase the size of the training data.
- To generate synthetic data :
 1. A system is trained in reverse direction i.e. source as **summary** and target as **text**.
 2. The reverse system is used to generate text for the given summary. Now, synthetic data is ready.
 3. Mix the real and synthetic data and train the final system.



Generation of synthetic data using reverse system.

Dataset

- **Real data (SwissText dataset)**

- **Synthetic data (Common Crawl)**

1. Build Vocabulary (using SwissText dataset, most frequent German words).
2. Select sentences based on the prepared Vocabulary. From the selected sentences, randomly choose 100K.
3. Generate synthetic data by using 100K sentences to input to the reverse trained model.

Dataset	#Text	#Summaries
Train	90K	90K
Dev	5K	5K
Test	5K	5K
Test Evaluation	2K	-

Statistics of experimental data (real) including the number of text and summaries.

Dataset	#Text	#Summaries
Train	190K	190K

Statistics of experimental data (real + synthetic) including the number of text and summaries.

Evaluation

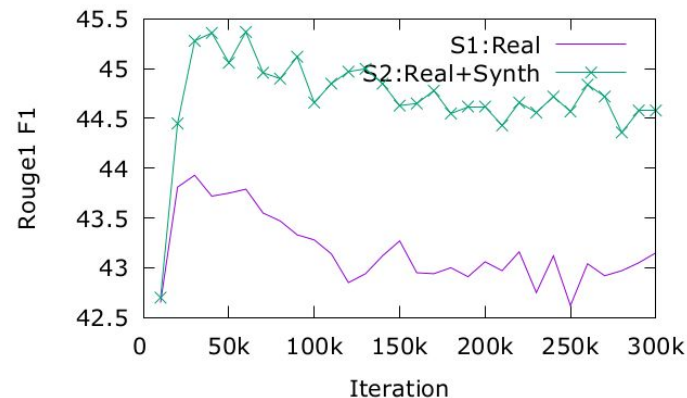
Setting	Dataset	Rouge_1_F1	Rouge_2_F1
S1	Dev	43.9	28.5
	Test	39.7	22.9
S2	Dev	45.4	29.8
	Test	55.7	41.8

Evaluation results of our models

Team	Rouge_1	Rouge_2
Shantipriya Parida, and Petr Motlicek (s2)	40.2	22.2
Dmitrii Aksenov, Georg Rehm, Julian Moreno Schneider	40.4	21.9
Nikola Nikolov	34.7	19.3
Valentin Venzin, Jan Deriu, Didier Orel, Mark Cieliebak	39.8	23.4
Pascal Fecht	40.9	23.5

SwissText 2019 Text Summarization Challenge Result

Source: http://ceur-ws.org/Vol-2458/summarization_challenge.pdf

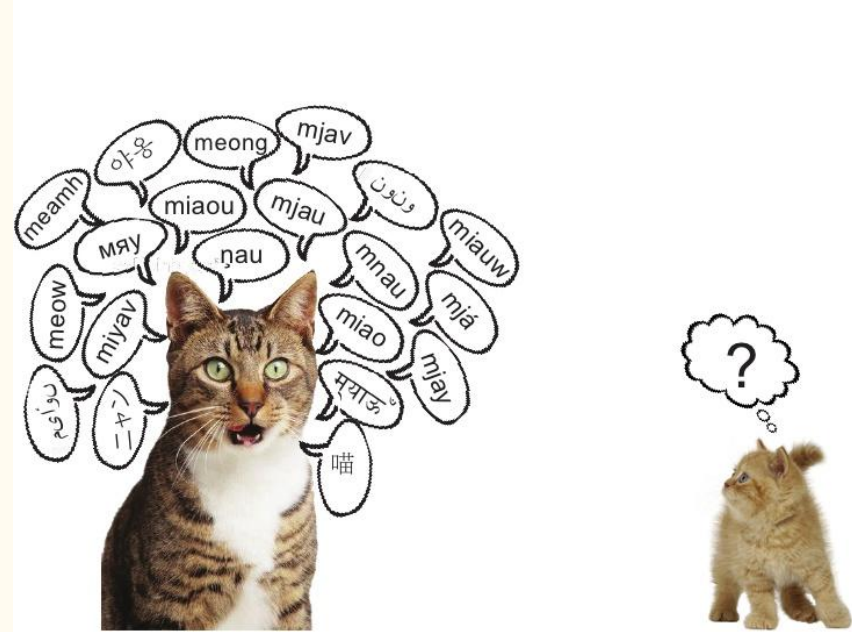


Learning curves in terms of Rouge 1 F1 Score on dev set

- Evaluations made using Rouge (Recall-Oriented Understudy for Gisting Evaluation) score, a popular metric for text summarization.

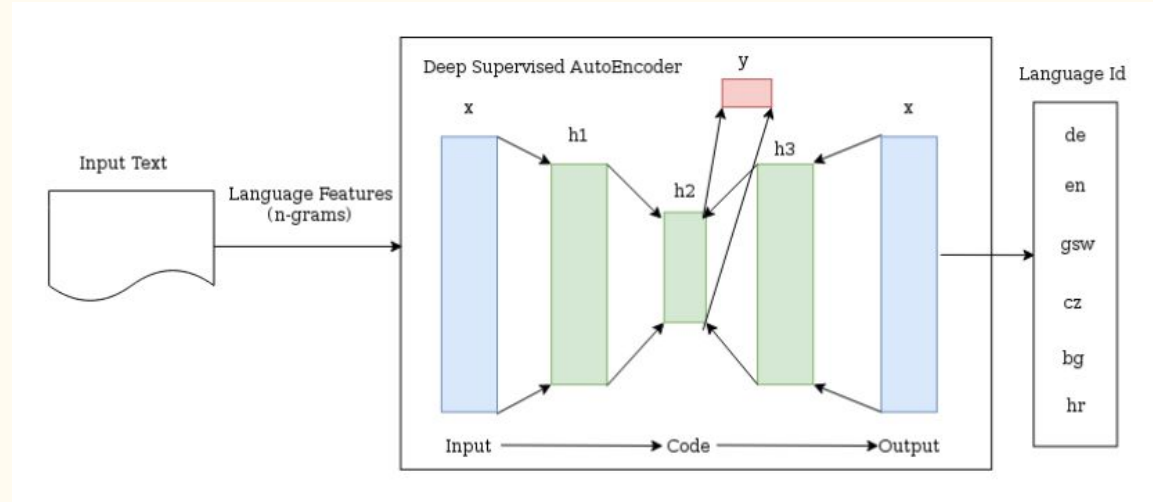
Case4: Language and Dialect Detection

- Its challenging to detect languages that have similar origins or dialects (e.g. German dialect identification, Indo-Aryan language identification)
- It may not be possible to distinguish related dialects with very similar phoneme and grapheme inventories for some languages.



Method Description

- We used character n-gram for extracting features from the input text.
- Extracted features are input to the deep supervised autoencoder (SAE).
- Bayesian optimizer used for selecting the optimal hyperparameters.



Proposed model architecture

Dataset

DSL Dataset: Discriminating between Similar Language (DSL) contains 13 different languages based on 6 different language group. We used DSLCCv2.0 in our experiment.

Ling10 Dataset: It contains 190,000 sentences categorized into 10 languages (*English, French, Portuguese, Chinese Mandarin, Russian, Hebrew, Polish, Japanese, Italian, Dutch*).

ILI Dataset: The Indo-Aryan Language Identification (ILI) dataset contains 5 closely-related languages of the Indo-Aryan language family – Hindi (also known as Khari Boli), Braj Bhasha, Awadhi, Bhojpuri, and Magahi.

Group Name	Language	Id
South Eastern Slavic	Bulgarian	bg
	Macedonian	mk
South Western Slavic	Bosnian	bs
	Croatian	hr
	Serbian	sr
West-Slavic	Czech	cz
	Slovak	sk
Ibero-Romance(Spanish)	Peninsular Spain	es-ES
	Argentinian Spanish	es-AR
Ibero-Romance(Portuguese)	Brazilian Portuguese	pt-BR
	European Portuguese	pt-PT
Astronesian	Indonesian	id
	Malay	my

DSL Language Group. Similar languages with their language code.

Result

Dataset	Training	Development	Test
DSL	252,000	28,000	14,000
Ling10	140,000	-	50,000
ILI	70,351	10,329	9,692

Dataset Statistics

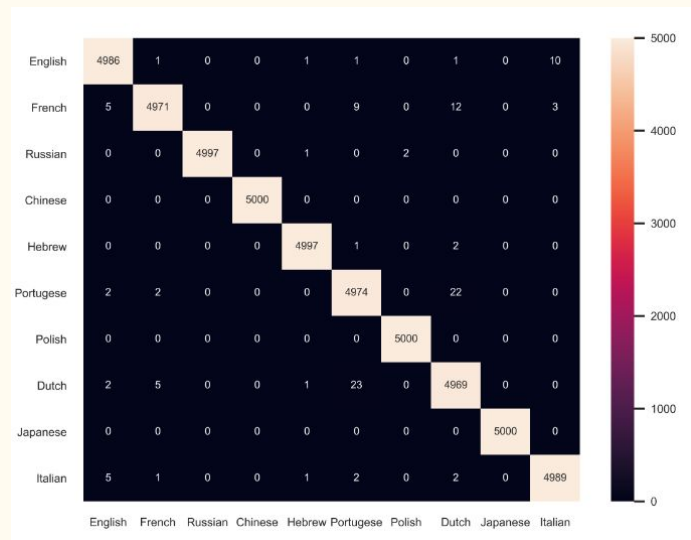
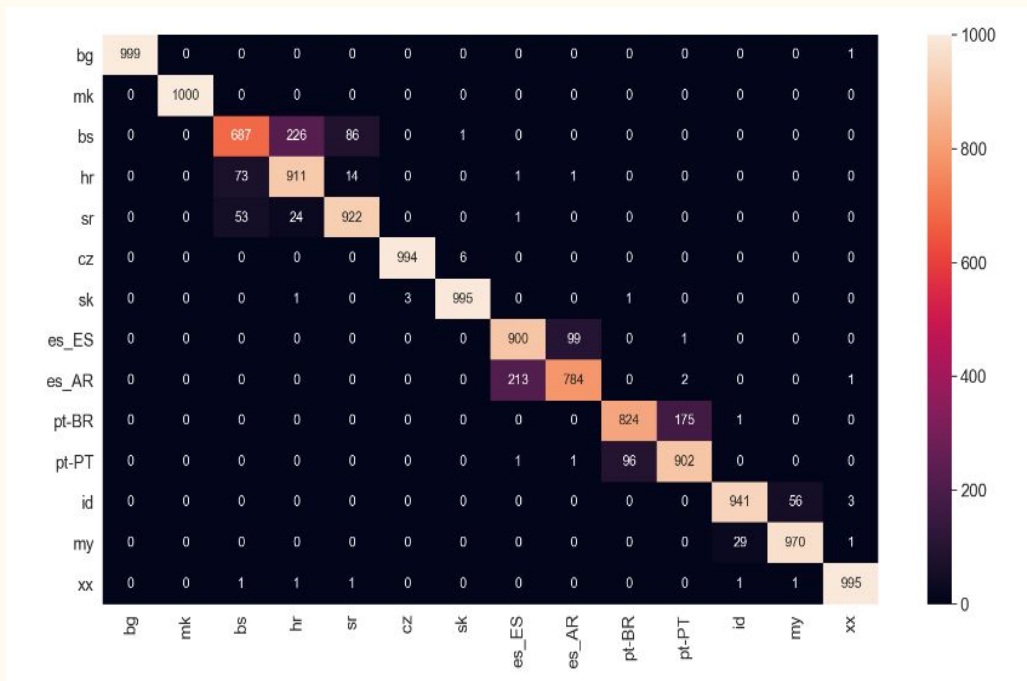
Model	Dataset	Accuracy (Test Set)
SAE (char-3gram)	Ling10	100%
SAE (char-3gram)	DSL	92%
SAE (char-3gram)	ILI	85%

Performance on test dataset.

Parameter	DSL	Ling10	ILI
<i>ngram</i> -range	1-3	1-3	1-3
number of target	14	10	5
embedding dimension	300	300	300
supervision	'clf'	'clf'	'clf'
converge threshold	0.00001	0.00001	0.00001
number of epochs	300	500	500

SAE model configurations for the dataset.

Result (Confusion Matrix)



Case 5: Fake News Detection @MEX-A3T



- The goal of IberLEF is to encourage the research community to organize competitive text processing, understanding and generation tasks in order to define new research challenges and setting new state-of-the-art results for the Natural Language Processing community, involving at least one of the following Iberian languages: Spanish, Portuguese, Catalan, Basque or Galician
- MEX-A3T 2020 had the following tracks:
 - Fake News detection
 - Aggressiveness detection
 - Both tracks contain documents in Mexican Spanish

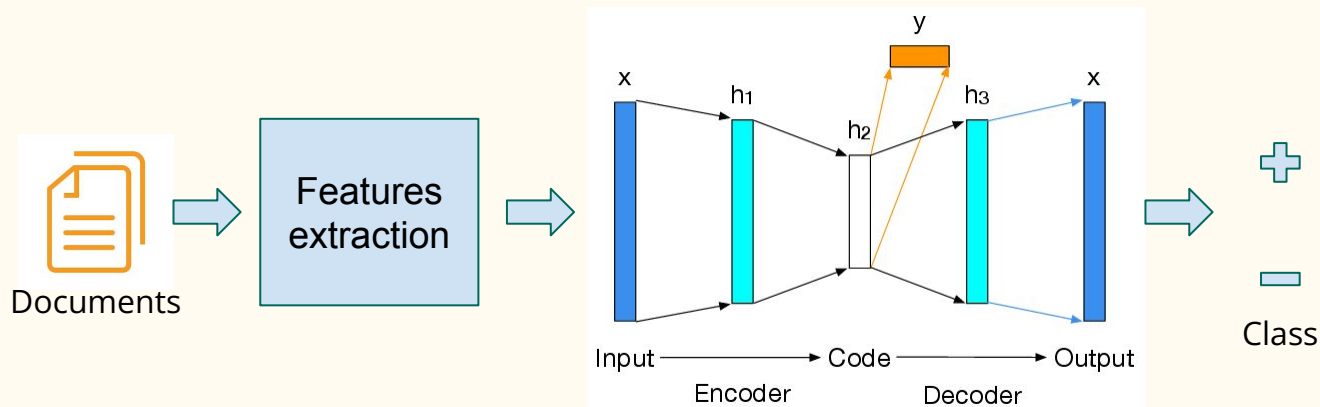
Fake News Detection



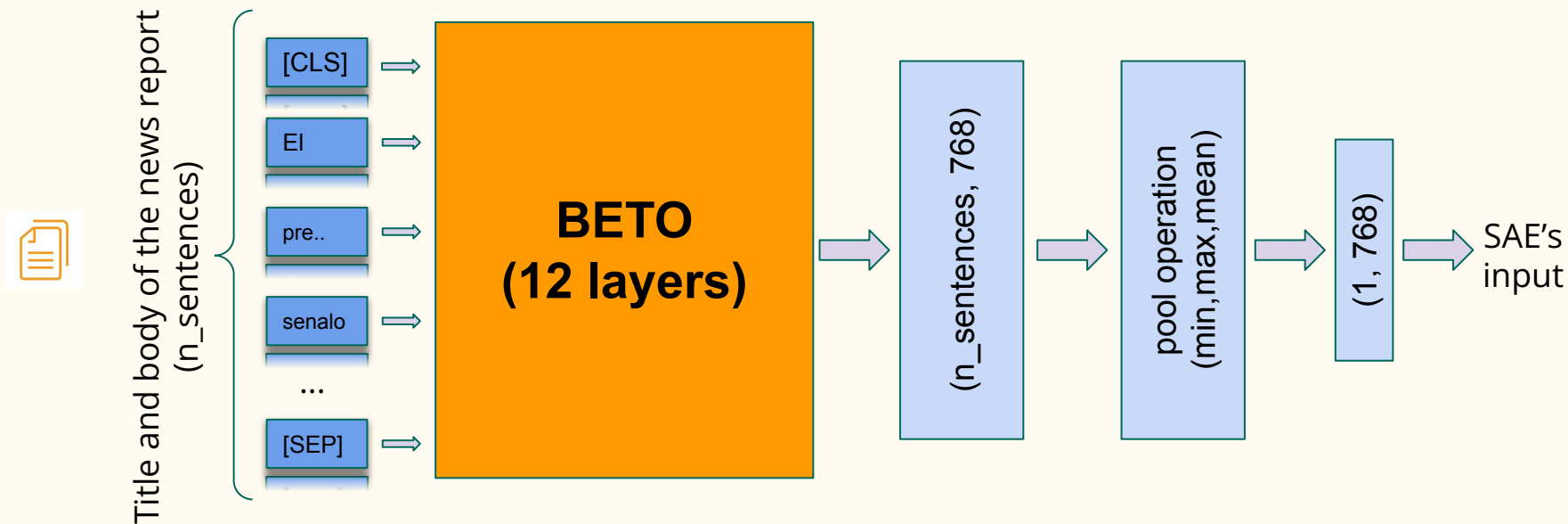
- Fake news provides information that aims to manipulate people for different purposes: terrorism, political elections, advertisement, satire, among others
- In social networks, misinformation extends in seconds among thousands of people
- A fake news detection system aims to help users detect and filter out potentially deceptive news
- The dataset consist of 971 documents, 676 for training and 295 for test
 - Documents are real news extracted from different news media in Mexico

Methodology

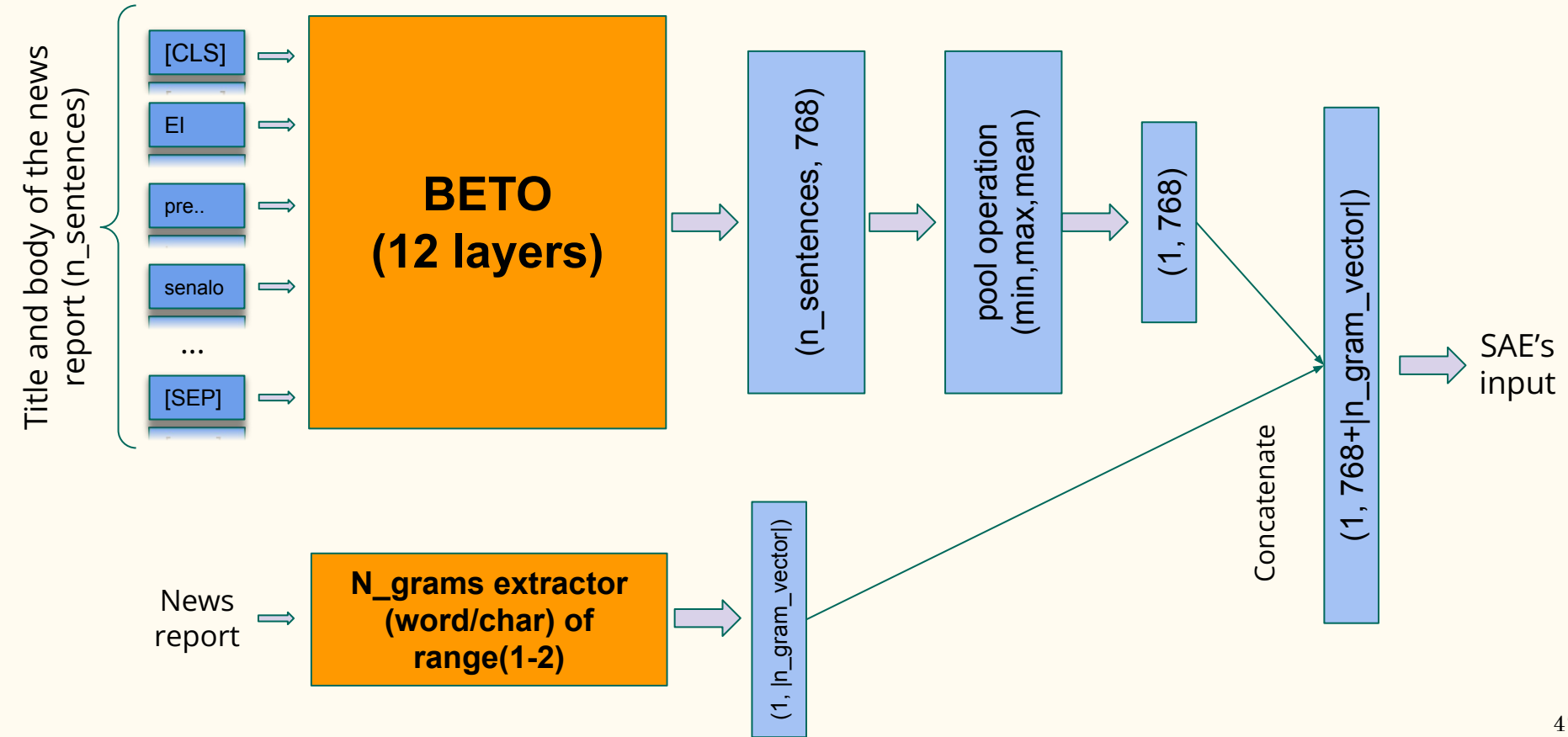
- Our goal was to evaluate the pertinence of deep SAE in these tasks
- As input features we used:
 - Spanish pre-trained **BERT** encodings (BETO)
 - Traditional text representation techniques such as **word** and **char n-grams** (ranges 1-2 and 1-3)
 - Combinations of BETO encodings plus traditional words/char n-grams vectors



Features extraction



Features extraction



Results (fake news task)

The best performance among 6 participating institutions

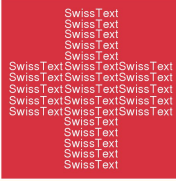


Table 3. Results in validation and test phases reported in F-score for real-news (F-), and macro average of F-score (Fm).

Input features	min-df,max-df	Validation phase			ID	Test phase		
		Fm	F+	F-		Fm	F+	F-
W(1,2)	0.01, 0.5	0.775	0.793	0.758	-	-	-	-
W(1,3)	0.01, 0.5	0.778	0.798	0.758	-	-	-	-
C(1,2)	0.01, 0.5	0.697	0.719	0.674	-	-	-	-
C(1, 3)	0.01, 0.5	0.757	0.768	0.745	-	-	-	-
B(min-pooling)		0.843	0.842	0.845	2	0.856	0.844	0.868
B(max-pooling)		0.830	0.830	0.830	-	-	-	-
B(mean-pooling)		0.833	0.831	0.835	-	-	-	-
C(1, 3)+W(1,2)	0.01, 0.5	0.805	0.807	0.802	-	-	-	-
B+W(1,2)	0.01, 0.3	0.845	0.846	0.844	1	0.850	0.840	0.859
B+C(1,3)	0.01, 0.3	0.834	0.834	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.3	0.833	0.831	0.835	-	-	-	-
B+W(1,2)+C(1,3)	0.01, 0.5	0.848	0.846	0.850	-	-	-	-
Third best system (in the track)						0.817	0.819	0.817
BOW-RF (baseline-given by track organizers)						0.786	0.785	0.787

Case 6: Operant Motive Detection

- According to Psycholinguistics theory, how we use language reveals information about our personality traits, educational level, age, etc.
- Operant methods are psychometrics, which are captured by having participants write free texts associated with faint images
 - Clinical research indicates that operant motives provide the possibility to assess behavioral long-term developments
- M - power, A - affiliation, L - achievement, F - freedom, O - zero, and corresponding levels (0 to 5).



Sample images that are shown to subjects during the OMT test

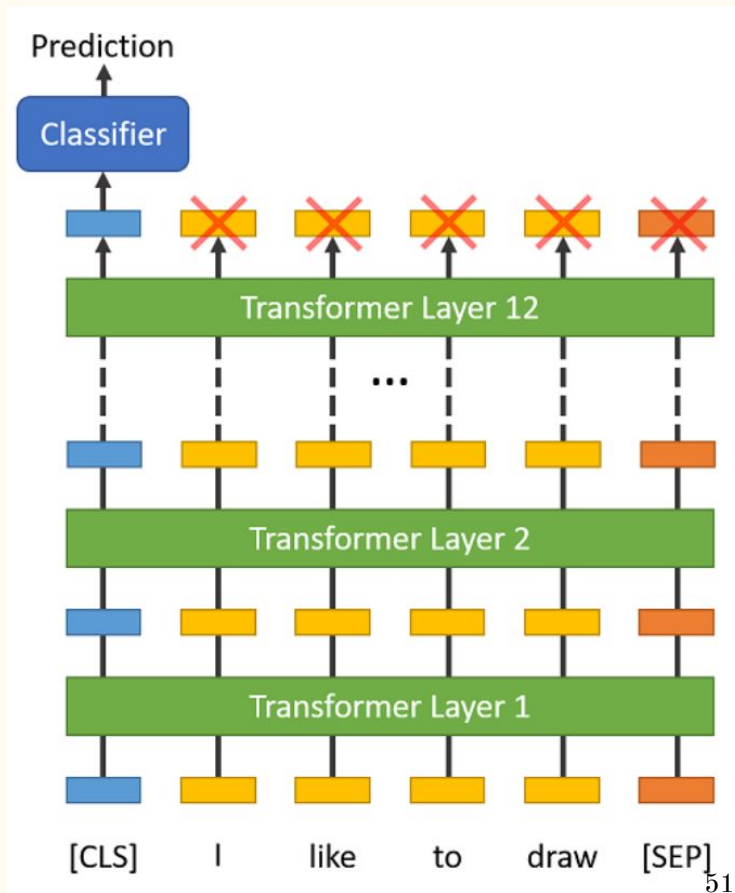
Task details

- The task is to predict motivational style solemnly based on tex
- The dataset:
 - Language: German
 - Training: 167,200*
 - Development: 20,900
 - Test: 20,900
 - Highly unbalanced

	Training	
	Average (σ)	Total
Tokens	20.27 (± 12.08)	3,389,945
Vocabulary	18.07 (± 9.82)	267,620
LR	0.92 (± 0.08)	0.08
	Development	
	Average (σ)	Total
Tokens	20.38 (± 12.17)	425,880
Vocabulary	18.17 (± 9.94)	55,606
LR	0.92 (± 0.08)	0.13
	Test	
	Average (σ)	Total
Tokens	20.24 (± 12.01)	423,018
Vocabulary	18.05 (± 9.76)	55,592
LR	0.92 (± 0.08)	0.13

Methodology^(1/3)

- **Simple transformers:** we add an untrained layer of neurons on the end, and re-train the model with the OMT classification task at the output
- **max_length** parameter is set to **90**, and models are re-trained up to **2 epochs**
- Three different configurations:
 - BERT
 - XLM
 - DistilBERT



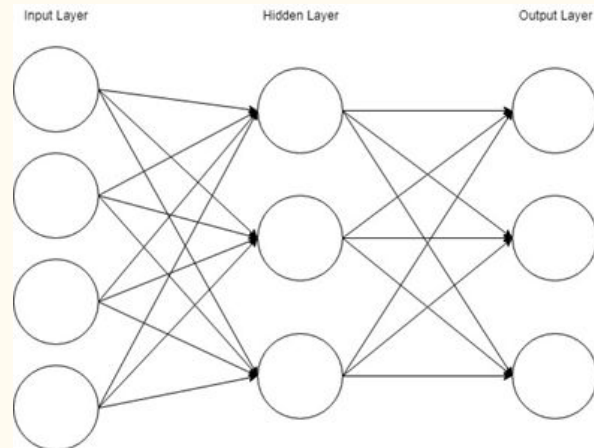
Methodology^(2/3)

- **Fully connected neural network (FC):** the FC is feed with the representation of the textual descriptions using:
 - **Pre-train BERT**
 - **Fine-tuned BERT**
- We reported results using two distinct ways for building the sentences representation

○ Last Hidden Layer

12 

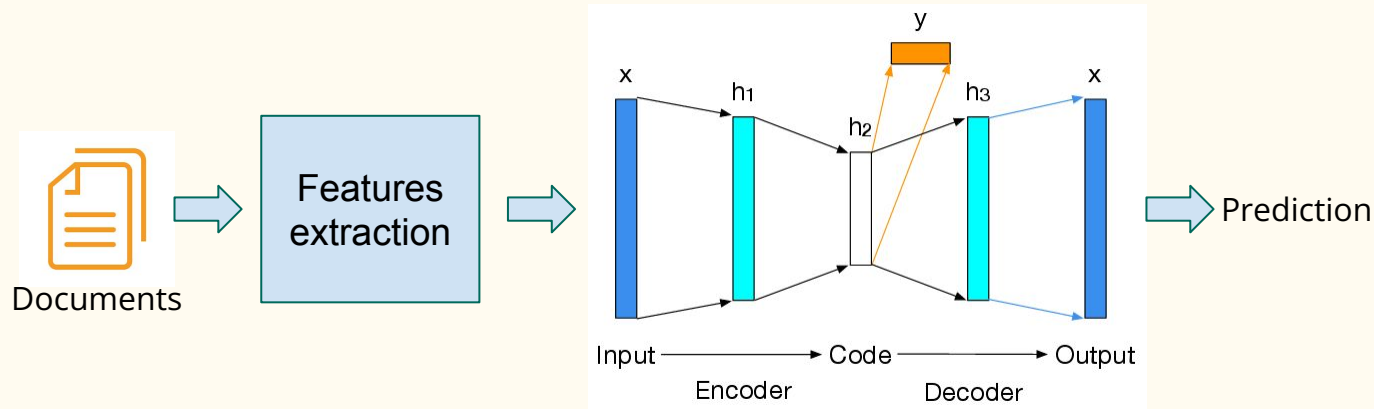
○ Concat Last
Four Hidden



Hyper Parameter	Range
number of layers	3
number of hidden layers	1
nodes in hidden layer	16
activation function	ReLU

Methodology^(3/3)

- We evaluate the performance of deep supervised autoencoders in the OMT task
- As input features we used:
 - German pre-trained and fine-tuned BERT encodings
 - Traditional text representation techniques such as word and char n-grams (ranges 1-2 and 1-3)
 - Combinations of BERT encodings plus traditional words/char n-grams vectors



Results (*test phase*)

The 2nd best performance among 3 (official) participating institutions



Method	Configuration type	Configuration sub-type	F1-macro (dev)	F1-macro (test)
ST	Bert	bert-base-german-cased	0.694	0.698
ST	XLM	xlm-mlm-ende-1024	0.688	0.686
ST	DistilBert	distilbert-base-german-cased	0.692	0.688
FC	Bert (pre-trained)	LHL	0.589	0.589
FC	Bert (pre-trained)	Concat4LHL	0.616	0.579
FC	Bert (fine-tuned)	LHL	0.673	0.671
FC	Bert (fine-tuned)	Concat4LHL	0.675	0.230
Baseline	SVM	<i>tf-idf</i>	0.639	0.644
1st place	—	—	—	0.704

Case 7 - Language Model Development

BertOdia

- Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited.
- We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.
- Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction.

BertOdia

Data and Model

- Building a language model is a challenging task in the case of low resource languages where the availability of contents is limited.
- We focus on building a general language model using the limited resources available in the low resource language which can be useful for many language and speech processing tasks.
- Our key contribution includes building a language-specific BERT model for this low resource Odia language and as per our best knowledge, this is the first work in this direction.

Source	Sentences	Unique Odia tokens
OdiEnCorp2.0	97,233	1,74,045
CVIT PIB	58,461	66,844
CVIT MKB	769	3,944
OSCAR	1,92,014	6,42,446
Wikipedia	82,255	2,36,377
Total Deduped	430,732	11,23,656

Table . Dataset statistics.

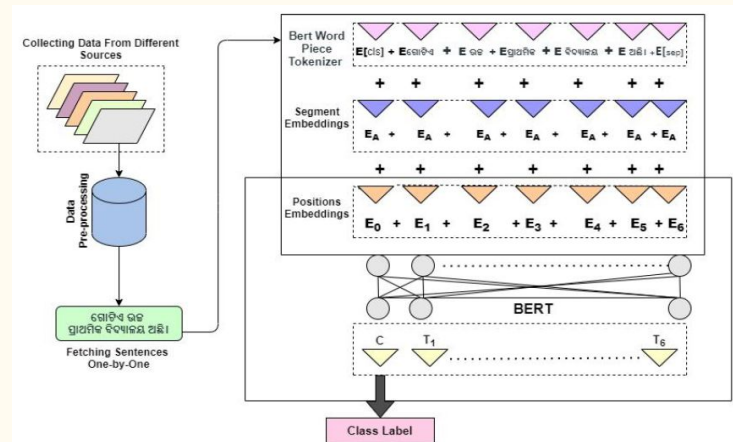


Fig: The Proposed Model: Visualisation of an our experimental model used for the Single Sentence Classification Task with Bert Embedding Layers.

BertOdia

BERT/RoBERTa Model Training

Parameter	BERT	RoBERTa
Learning Rate	5e-5	5e-5
Training Epochs	5	10
Dropuout Prob	0.1	0.1
MLM Prob	0.1	0.2
Self attention layer	6	6
Attention head	12	12
Hidden layer size	768	768
Hidden layer Activation	gelu	gelu
Total parameters	84M	84M

Table 2. Training Configurations

Model	Text Classification Accuracy
BertOdia	96.0
RoBERTaOdia	92.0
ULMFit	91.9

Table 3. BertOdia Performance

BertOdia

IndicGlue Task

- For the Cloze-style Multiple-choice QA task, we feed the masked text segment as input to the model and we fine-tune the model using cross-entropy loss.
- For the Article Genre Classification task we used the IndicGLUE dataset for news classification.

Model	Article Genre Classification	Cloze-Style multiple-choice QA
XLM-R	97.07	35.98
mBERT	69.33	26.37
IndicBERT base	97.33	39.32
IndicBERT large	97.60	33.81
BertOdia	96.90	23.00

Table: Comparison of BertOdia with IndicBERT. BertOdia was trained on 6% of the data of IndicBERT.

- The code and dataset are available at:

https://colab.research.google.com/gist/satyapb2002/aeb7bf9a686a9c7294ec5725ff53fa49/odiabert_language_model.ipynb

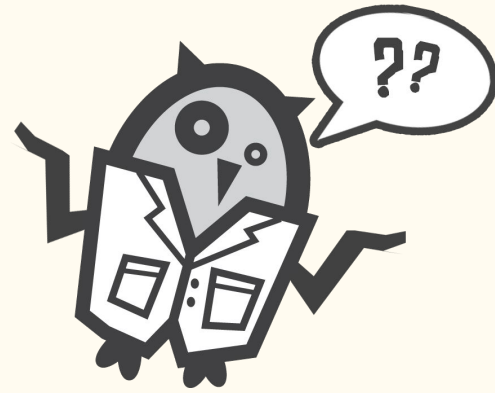
Summary

- SAE with Bayesian Optimization for the language detection task found effectively for discriminating between very close languages or dialects
- SAE are able to generalize well, however, they seem to perform better on texts extracted from formal written
 - Fake news detection, best performance, documents extracted from real news media
- SAE are less computationally expensive as compared to attention based DL models (e.g., transformers)
 - They do not require high volume of data

Q&A

Contact information:

- Twitter: @Shantipriyapar3
- Web : shantipriya.me
- Email: shantipriya.parida@gmail.com



Thank You

