The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI 2021)

Simple or Complex? Learning to Predict Readability of Bengali Texts

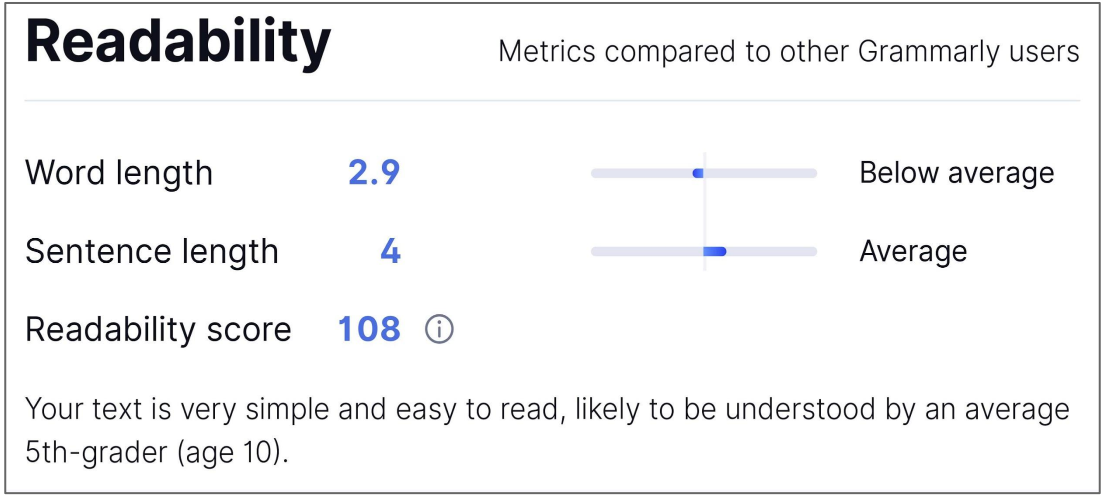
Susmoy Chakraborty[1*], Mir Tafseer Nayeem[1*], Wasi Uddin Ahmad[2]

[1]Ahsanullah University of Science and Technology

[2]University of California, Los Angeles

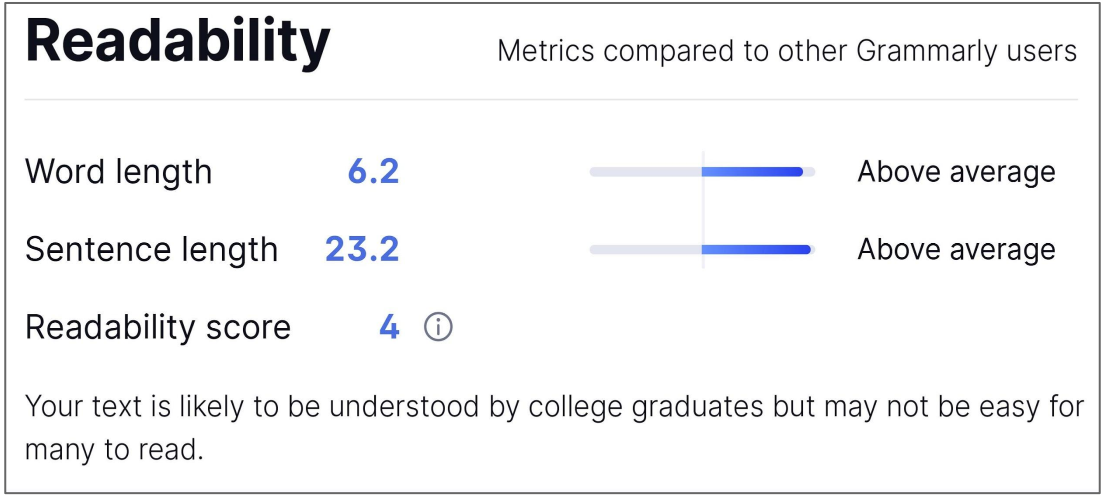
[*]**Equal Contribution, listed by alphabetical order**

What is Readability?



Readability measuring of a document using **Grammarly**

What is Readability?



Readability measuring of a document using **Grammarly**

What is Readability?

Measures how much energy the reader will have to expend in order to understand a writing at optimal speed and find interesting

What is Readability?

Measures how much energy the reader will have to expend in order to understand a writing at optimal speed and find interesting

First step of Text Simplification

Formulas for measuring Readability

- Automated Readability Index (Senter and Smith 1967)
- Flesch reading ease (Flesch 1948)
- Flesch–Kincaid grade level (Kincaid et al. 1975)
- Gunning Fog index (Gunning 1952)
- SMOG (Mc Laughlin 1969)
- Dale–Chall formula (Dale and Chall 1948, Chall and Dale 1995)

Output: A score that estimates the grade level or years of education of a reader based on the **U.S education system**

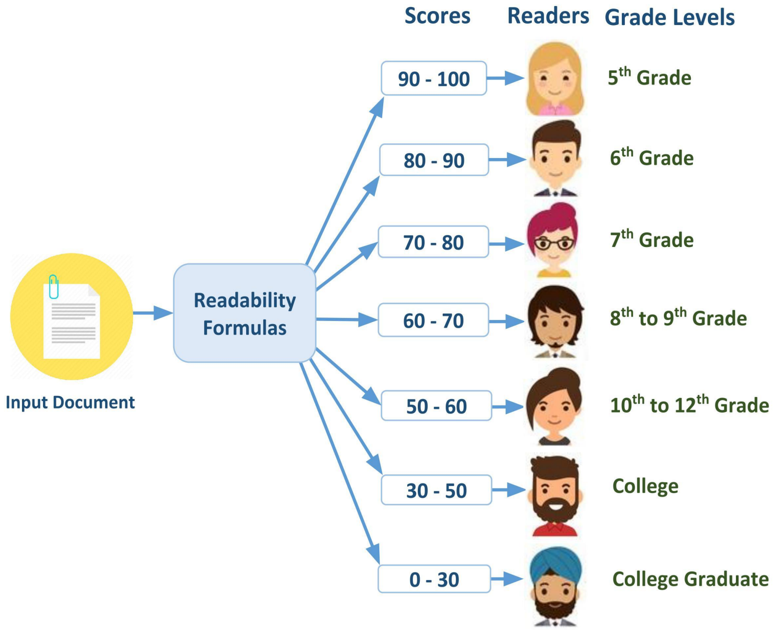
Output: Generally Correlate highly with the actual readability of an English text

Formulas for measuring Readability

- Automated Readability Index (Senter and Smith 1967)
- Flesch reading ease (Flesch 1948)
- Flesch–Kincaid grade level (Kincaid et al. 1975)
- Gunning Fog index (Gunning 1952)
- SMOG (Mc Laughlin 1969)
- Dale–Chall formula (Dale and Chall 1948, Chall and Dale 1995)

These formulas are still used by commercial readability measuring tools such as **Grammarly** and **Readable**

Formulas for measuring Readability: Visual representation



Readability prediction task

Formulas for measuring Readability: Features

Average Sentence Length (**#words** / **#sentences**)
Average Word Length (**#characters** / **#words**)
Number of Syllables
Number of Difficult words

And so on...

Responsible for **simplicity** or **complexity** of an English document

Fields where Readability measurement is used

Education

Government

Health care

Websites

Dyslexia

Readability formulas on **non-English** texts

Readability formulas on **non-English** texts

Not Straightforward like English**!**

Readability formulas on **non-English** texts

Not Straightforward like English!

Are all the readability measuring formulas **language-independent**?

Example: 3000 easy English words list for the Dale–Chall formula

Readability formulas on **non-English** texts

Not Straightforward like English!

Are all the readability measuring formulas **resource-independent**?

Readability formulas on **non-English** texts

Not Straightforward like English!

Are all the readability measuring formulas **resource-independent**?

Resources, e.g., Syllable counting tool, stemmer, lemmatizer are required for readability measuring formulas

Readability formulas on **non-English** texts

Not Straightforward like English!

Are all the readability measuring formulas **resource-independent**?

Resources, e.g., Syllable counting tool, stemmer, lemmatizer are required for readability measuring formulas

Obstacle for the readability analysis of **low-resource-languages** (e.g., **Bengali**)

Related Works: Non-English languages (except Bengali)

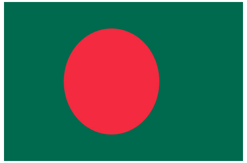
Japanese: Sato 2014
Russian: Reynolds 2016
French: Seretan 2012
Swedish: Grigonyte et al. 2014
Polish: Broda et al. 2014
Arabic: El-Haj and Rayson 2016
Vietnamese: Nguyen and Uitdenbogerd 2019
German: Battisti et al. 2020

Readability analysis tool

Arabic (Al-Twairesh et al. 2016), Italian (Okinina, Frey, and Weiss 2020), Japanese (Sato, Matsuyoshi, and Kondoh 2008)

Selecting **Bengali** language for our non-English Readability research

Native language of **Bangladesh**, also used in **India** (e.g., West Bengal, Tripura)



7th most spoken language in the world, **250 million** native speakers[1]

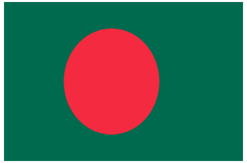
Suffers from a lack of **fundamental resources** for Natural Language Processing (NLP)

[1] https://w.wiki/57

Selecting **Bengali** language for our non-English Readability research

Suffers from a lack of **fundamental resources** for Natural Language Processing (NLP)

For example, no spoken syllable counter available for the Bengali language, where **syllable count** feature is widely used in traditional readability formulas



Related Works: **Bengali** language

- Das and Roychoudhury 2006
- Islam, Mehler, and Rahman 2012
- Sinha et al. 2012
- Islam, Rahman, and Mehler 2014
- Phani, Lahiri, and Biswas 2014
- Sinha and Basu 2016
- Phani, Lahiri, and Biswas 2019

Summary of previous Bengali Readability research works

- **Dataset:** Bengali textbook (Bangladeshi), literature, etc.
- Traditional readability formulas were applied to Bengali dataset by Islam, Mehler, and Rahman 2012; Islam, Rahman, and Mehler 2014; Sinha et al. 2012

Related Works: **Bengali** language

Summary of previous Bengali Readability research works

- Some of these works developed new formulas/models using Regression Analysis (e.g., Sinha et al. 2012; Phani, Lahiri, and Biswas 2019)
  - ➢ Various features extracted from Bengali documents, **significant features**: **Average Sentence length**, **Consonant Conjunct**, etc.
- Machine Learning methods (SVM, SVR) used by Sinha and Basu 2016

Research Objective

Are **previous Bengali readability analysis works** satisfactory?

These works are **narrow** and sometimes **incorrect**!

Research Objective

Are **previous Bengali readability analysis works** satisfactory?

These works are **narrow** and sometimes **incorrect**!

Small scale dataset, **not publicly available**!

Research Objective

Are **previous Bengali readability analysis works** satisfactory?

These works are **narrow** and sometimes **incorrect**!

Small scale dataset, **not publicly available**!

In some cases, **unclear methodologies**!

Research Objective

Are **previous Bengali readability analysis works** satisfactory?

These works are **narrow** and sometimes **incorrect**!

Small scale dataset, **not publicly available**!

In some cases, **unclear methodologies**!

Importance of the feature **Consonant Conjunct** has been showed, but no specific algorithm found

Research Objective

Previous Bengali readability analysis works are narrow and sometimes <u>incorrect</u>!

Not straightforward to adapt readability formulas used for the English language

- These formulas (e.g., Automated Readability index) are developed for U.S. based education system
- Predict U.S grade level of the reader

Research Objective

Previous Bengali readability analysis works are narrow and sometimes incorrect!

Straightforward procedure is incorrect for the Bengali language, but why?

Because Bangladeshi education system and grade level[2] are **different** from U.S!

So, in the case of previous Bengali readability works, grade level mapping is **faulty** and led to **incorrect results**

[2] https://www.scholaro.com/pro/Countries/bangladesh/Education-System

Research Objective

Previous Bengali readability analysis works are narrow and sometimes incorrect!

Straightforward procedure is incorrect for the Bengali language, but why?

Because Bangladeshi education system and grade level[2] are **different** from U.S!

So, in the case of previous Bengali readability works, grade level mapping is **faulty** and led to **incorrect results**

How can we solve this problem? Please see in the next slide!

[2] https://www.scholaro.com/pro/Countries/bangladesh/Education-System

Research Objective

Strong relationship between **reading skills** and **human cognition**, which varies depending on **different age groups** (Riddle 2007)

⬇

In this work, we map grade level to different age groups to present

age-to-age comparison

Previous work: Grade level comparison of Bangladeshi and U.S. education systems

Our work: Age-to-age comparison of Bangladeshi and U.S. education systems

Research Objective: Our main **Contributions**

- We correctly adapt document-level readability formulas traditionally used for U.S. based education system to the Bengali education system with a **proper age-to-age comparison**.
- A document level dataset consisting of **618** documents with **12 different grade levels** for the evaluation of traditional readability formulas.
- An **efficient algorithm for counting consonant conjuncts** from a given word, with a human annotated corpus comprising 341 words for evaluating the effectiveness of this algorithm.

Research Objective: Our main **Contributions**

- We further divide the document-level task into **sentence-level** due to the long-range dependencies of RNNs and the unavailability of large scale human annotated corpora.
  - ➢ **96,335** sentences with **simple** and **complex** labels to experiment with supervised neural models
  - ➢ We design neural architectures and use **all available pretrained language models** of the Bengali language
  - ➢ These neural architectures will serve as a **baseline** for future Bengali readability prediction works

Research Objective: Our main **Contributions**

- These resources can be helpful for **several other tasks**!
- We Design a **Bengali readability analysis tool**, which would be useful for educators, content writers or editors, researchers, and readers of different ages

Dataset

Documents from several published textbooks, popular sources from **Bangladesh** and **India**

- **Most common** and **very well-known** among children and adults
- Usually published after rigorous review and editorial process, **widely read by various age groups**

Dataset

Documents from several published textbooks, popular sources from **Bangladesh** and **India**

- ➢ **Most common** and **very well-known** among children and adults
- ➢ Usually published after rigorous review and editorial process, **widely read by various age groups**

In this work, for readability prediction we present two datasets

- **Document-level dataset** to experiment with formula-based approaches
- **Sentence-level dataset** to train supervised neural models

Methodology

Formula-based Approaches

Document-level dataset

NCTB

16 Textbooks from **class 1 to 12** provided by National Curriculum and Textbook Board (NCTB), Bangladesh[3]

Additional Sources

Documents (Literature and articles) from various popular and well known sources for both children and adults

Dataset	#Docs	Avg. #sents	Avg. #words
NCTB	380	66.8	585.8
Additional	238	391.2	3045.0

Statistics of the Document-level dataset

618 Documents

[3] https://w.wiki/ZwJ

Formulas-based Approaches: **Experiment**

In this work, we use 6 readability formulas:

- Automated Readability Index (ARI)
- Flesch reading Ease (FE)
- Flesch–Kincaid (FK)
- Gunning Fog (GF)
- SMOG
- Dale–Chall (DC)

Number of Documents: **14** (10 from NCTB, 4 from Additional) from Document-level dataset

Only 14 out of 618 documents!

But why?

Formulas-based Approaches: **Experiment**

In this work, we use 6 readability formulas:

- Automated Readability Index (ARI)
- Flesch reading Ease (FE)
- Flesch–Kincaid (FK)
- Gunning Fog (GF)
- SMOG
- Dale–Chall (DC)

Only 14 out of 618 documents,
but why?

Because of the unavailability of **spoken syllable counting system** for the Bengali language

Three formulas require a common feature, which is **the number of syllables**

Formulas-based Approaches: **Experiment**

- We use a pronunciation dictionary[4] for the Bengali language with more than 67k words provided by **Google Language Resources** as our **syllable count dictionary**
- **ARI: Language Independent**, no need of extra resources!
- **DC:** We manually annotate **3,396 Bengali easy words** (based on the word frequency of children type documents) as an alternative to 3,000 easy English words list

[4] https://git.io/JJhdm

Formulas-based Approaches: **Performance**

Bangladeshi education system

Usually, children are admitted to class 1 at the age of 6, and complete higher secondary education (Class 12) at the age of 17[5]

[5] https://www.scholaro.com/pro/Countries/bangladesh/ Education-System

Formulas-based Approaches: **Performance**

Document	BN age	ARI	U.S. age	FE	U.S. age	FK	U.S. age	GF	U.S. age	SMOG	U.S. age	DC	U.S. age
Class 1	6	1	**5-6**	40.9	18-22	9	14-15	6	11-12	N/A	-	5.9	10-12
Class 2	7	1	5-6	30.6	18-22	10	15-16	10	15-16	9	14-15	5.3	10-12
Class 3	8	3	**7-9**	21.9	≥21	12	17-18	11	16-17	10	15-16	7.2	14-16
Class 4	9	3	**7-9**	34.1	18-22	10	15-16	9	14-15	9	14-15	7.3	14-16
Class 5	10	6	11-12	11.0	≥21	13	18-19	15	20-21	12	17-18	7.4	14-16
Class 6	11	4	9-10	21.1	≥21	12	17-18	14	19-20	11	16-17	8.2	16-18
Class 7	12	6	**11-12**	13.1	≥21	13	18-19	13	18-19	11	16-17	7.2	14-16
Class 8	13	6	11-12	16.2	≥21	13	18-19	13	18-19	12	17-18	8.5	16-18
Class 9/10	14-15	12	17-18	-8.6	-	18	≥20	20	≥21	17	≥19-20	7.3	**14-16**
Class 11/12	16-17	11	**16-17**	-2.6	-	18	≥20	19	≥21	16	≥19-20	8.1	**16-18**
Children 1	6-10	1	**5-6**	32.0	18-22	10	15-16	8	13-14	8	13-14	5.0	**10-12**
Children 2	6-10	2	**6-7**	33.8	18-22	10	15-16	9	14-15	9	14-15	6.1	12-14
Adults 1	≥18	12	**17-18**	-22.8	-	21	**≥20**	24	**≥21**	19	**≥19-20**	11.5	**≥21**
Adults 2	≥18	3	7-9	27.3	**≥21**	11	16-17	10	15-16	9	14-15	7.1	14-16

Formulas-based approaches: **Limitation**

Some of these formulas depend on the number of words or number of sentences.

- SMOG: At least 30 sentences!
- Gunning Fog: At least 100 words!

We tackle this problem in our **Supervised Neural Approaches**

Methodology

Supervised Neural Approaches

Supervised Neural Approaches

We divide the document-level task into a **supervised binary sentence classification problem**

➢ Classes: **Simple** and **Complex**

Why we convert Document-level task into sentence-level task?

Supervised Neural Approaches

We divide the document-level task into a **supervised binary sentence classification problem**

- ➢ Classes: **Simple** and **Complex**

Why we convert Document-level task into sentence-level task?

- Document-level understanding is challenging, **insufficient Document-level dataset**
- Long-range dependencies of RNNs (Truinh et al. 2018)

Sentence-level Dataset

We break documents from **Document-level Dataset** (NCTB + Additional) into sentences to create a large-scale dataset for training neural models

Simple Documents	**Complex Documents**
Class 1 to 5 (6 to 10 years old students) from **NCTB**, all children type documents from **Additional**	No documents from **NCTB**, all adult type documents from **Additional**
Sentences from these documents are labeled as **Simple**	Sentences from these documents are labeled as **Complex**

Sentence-level Dataset

- Some simple sentences exist in complex sentences, we remove these using semantic similarity (fastText pretrained model for the Bengali language, Grave et al. 2018)

NOTE

Sentences from Sentence-level Dataset are editor-verified and further annotated by us

	Train	Dev	Test
Simple Sentences			
#Sents	37,902	1,100	1,100
Avg. #words	8.16	7.97	8.31
Avg. #chars	44.71	43.85	45.57
Complex Sentences			
#Sents	54,033	1,100	1,100
Avg. #words	8.04	8.08	8.16
Avg. #chars	44.01	44.65	44.63

Statistics of the Sentence-level dataset

Supervised Neural Approaches: **Additional Feature Fusion**

- **C**haracter **L**ength (**CL**): Total number of characters in a sentence including white spaces

- **C**onsonant **C**onjunct (**CC**): Total number of consonant conjuncts in a sentence

Simple: আমরা এই সব পোশাক প্রতিদিন পরি
[We wear all these clothes everyday]
CL: 30
CC: আমরা এই সব পোশাক প্রতিদিন পরি **= 1**

Complex: তাহার ওষ্ঠাধরের উভয় প্রান্ত ঈষৎ প্রসারিত হইল মাত্র
[Only the ends of his lips were slightly extended]
CL: 50
CC: তাহার ওষ্ঠাধরের উভয় প্রান্ত ঈষৎ প্রসারিত হইল মাত্র **= 5**

Visual representation of CL and CC for a **Simple** and a **Complex** sentence

Supervised Neural Approaches: **Additional Feature Fusion**

To evaluate this **CC count algorithm**, we manually create **a dataset** with 341 words and their corresponding CC count

- ➤ **Performance: 100%** accuracy has been achieved!

Algorithm 1: Consonant conjunct count algorithm.

```
Procedure ConsonantconjunctCount (W)
    Data: Input word W, which is an array of Bengali
          characters.
    Result: Return the number of consonant conjuncts in
            input word W.
    A ← Bengali sign VIRAMA (Wikipedia 2020);
    cc_count ← 0;
    l ← length(W);
    for k ← 0 to l − 1 do
        if W[k] == A then
            if k − 1 ≥ 0 and k + 1 < l then
                if k − 2 ≥ 0 then
                    if W[k − 1] and W[k + 1] is a
                      Bengali Consonant and W[k − 2]
                      != A then
                        cc_count ← cc_count + 1;
                    end
                end
                else if W[k − 1] and W[k + 1] is a
                  Bengali Consonant then
                    cc_count ← cc_count + 1;
                end
            end
        end
    end
    return cc_count;
```

Supervised Neural Approaches: **Ablation Experiment**

- **Baseline Models:** Bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997), BiLSTM with attention mechanism (Raffel and Ellis 2016)

- We extend BiLSTM model by adding **Global Average Pooling** and **Global Max Pooling** (Boureau, Ponce, and LeCun 2010)

Supervised Neural Approaches: **Ablation Experiment**

- **Baseline Models:** Bidirectional LSTM (BiLSTM) (Schuster and Paliwal 1997), BiLSTM with attention mechanism (Raffel and Ellis 2016)

- We extend BiLSTM model by adding **Global Average Pooling** and **Global Max Pooling** (Boureau, Ponce, and LeCun 2010)

- **Ablation study:** We use this extended model to demonstrate the effects of CL and CC feature fusion

Supervised Neural Approaches: **Ablation Experiment**

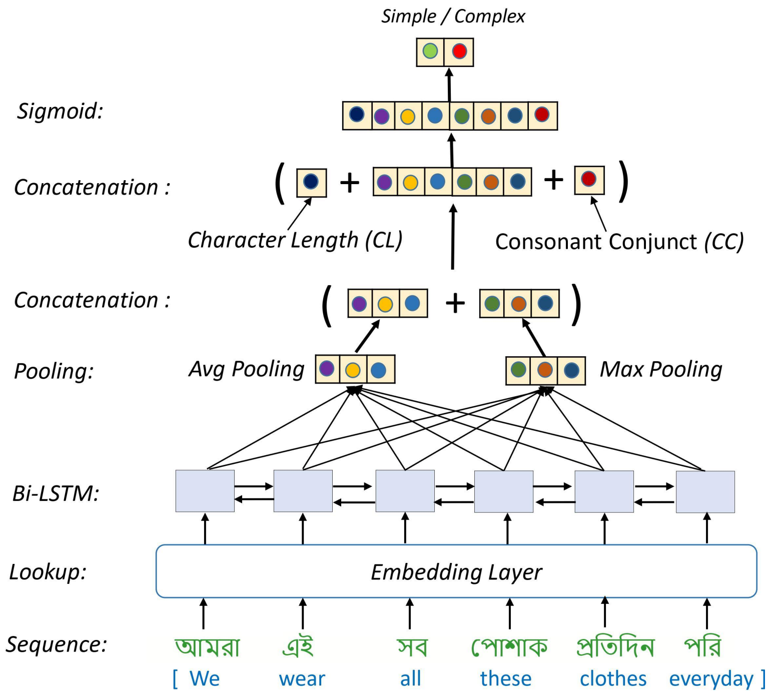
- We use **all pretrained language models** available to date for the Bengali language
  - ↪️ Word2vec (Mikolov et al. 2013)
  - ↪️ GloVe (Pennington, Socher, and Manning 2014)
  - ↪️ fastText (Grave et al. 2018)
  - ↪️ BPEmb (Heinzerling and Strube 2018)
  - ↪️ ULMFiT (Howard and Ruder 2018) provided by iNLTK[6]
  - ↪️ TransformerXL (Dai et al. 2019) provided by iNLTK[6]
  - ↪️ laserembeddings[7], which is based on LASER (Artetxe and Schwenk 2019)
  - ↪️ LaBSE (Feng et al. 2020): Language agnostic BERT

[6] https://git.io/JUItc

[7] https://pypi.org/project/laserembeddings/

Supervised Neural Approaches: **Ablation Experiment**

For each input sentence, we calculate **CL** and **CC** to concatenate with the pooling layers



Readability Prediction Model

Supervised Neural Approaches: **Performance**

Baseline Models				
Models	**A**	**R**	**P**	**F1**
BiLSTM	77.5	69.4	82.8	75.5
BiLSTM + Attention	76.4	65.9	83.3	73.6
Ablations				
Models	**A**	**R**	**P**	**F1**
BiLSTM with Pooling	81.3	78.8	83.0	80.8
+ Word2vec	85.5	80.2	89.7	84.7
+ CL + CC	85.7	80.9	89.5	85.0
+ GloVe	86.1	79.3	**91.9**	85.1
+ CL + CC	86.1	81.3	89.9	85.4
+ fastText	86.0	80.1	90.8	85.1
+ CL + CC	**86.4**	82.9	89.1	85.9
+ BPEmb	86.2	81.5	90.0	85.6
+ CL + CC	86.0	81.2	89.8	85.3
+ ULMFiT	85.5	77.6	**92.0**	84.2
+ CL + CC	86.2	80.4	91.0	85.4
+ TransformerXL	86.3	82.7	89.0	85.8
+ CL + CC	**86.7**	83.5	89.3	**86.3**
+ LASER	**86.4**	84.3	88.0	86.1
+ CL + CC	86.3	**84.6**	87.6	86.1
+ LaBSE	86.0	80.3	90.8	85.2
+ CL + CC	**86.7**	**86.5**	86.8	**86.7**

Supervised Pretraining

FastText supervised text classification techniques

Joulin et al. 2017

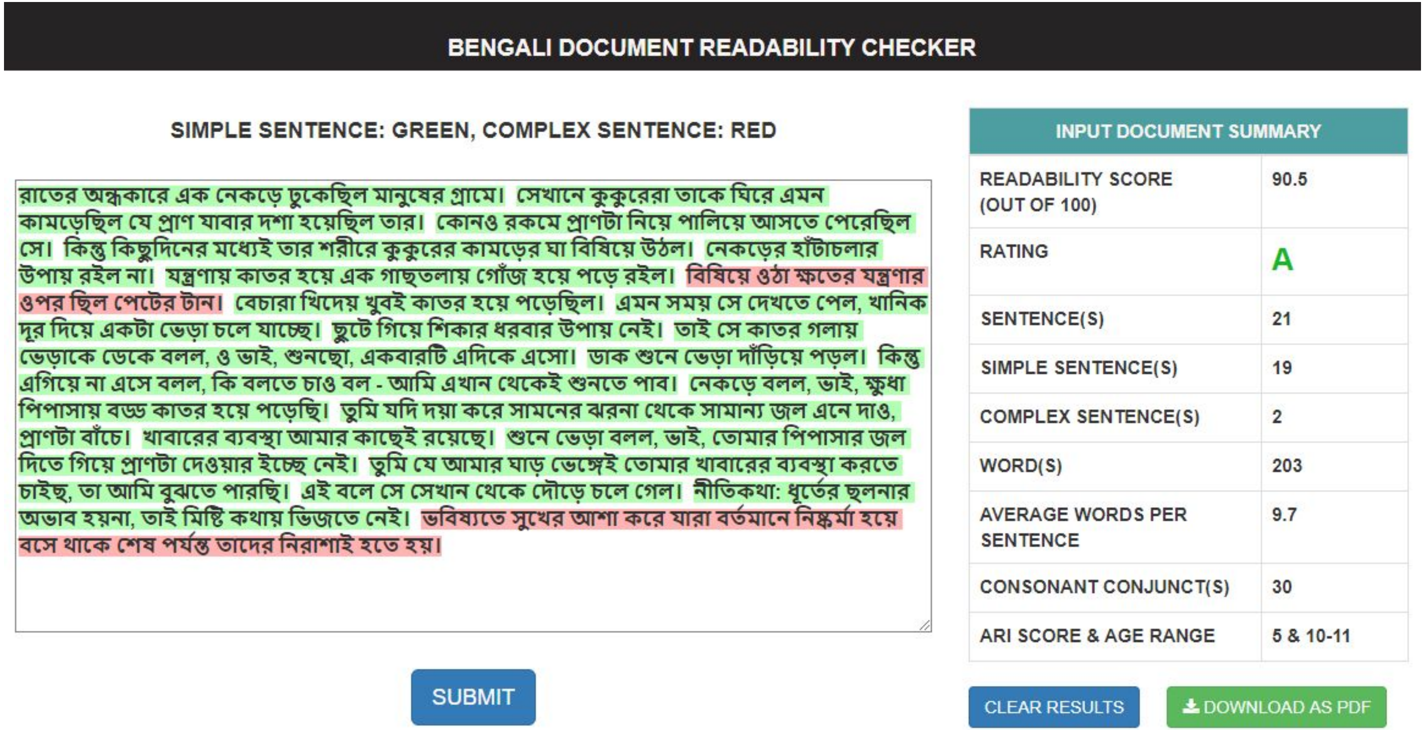
3 classifiers using word n-grams (unigram, bigram, trigram) and character n-grams (2 to 6 length)

Models	A	R	P	F1
fastText Unigram	86.0	82.8	88.4	85.5
fastText Bigram	86.6	84.9	87.9	86.4
fastText Trigram	**87.4**	**85.0**	**89.2**	**87.1**

Performance of Supervised Pretraining

Demo Video:
https://youtu.be/U05Pf9Y4tCQ

Bengali Readability Analysis Tool



Future Works

- Increasing sentence-level dataset
- Our tool-based user study
- Readability prediction of Bengali-English code-mixed texts

Our code, data and all other resources:

https://github.com/tafseer-nayeem/BengaliReadability

Thank You!

Questions?

You can also mail us at

susmoyaust36@gmail.com || mir.nayeem@alumni.uleth.ca || wasiahmad@ucla.edu