Vidya.AI Pitch Deck

March 6, 2023, | Chicago, Illinois

Ananth Prayaga
Garima Sohi
Kshitij Mittal
Manish Kumar
Urvaj Shah

Vidya

vid•ya (vid′yä)

Vidya is a Sanskrit noun meaning "right knowledge" and "clarity"
It is frequently used in South Asia, implying the conception of knowledge and learning.

Business Overview

Problem Statement

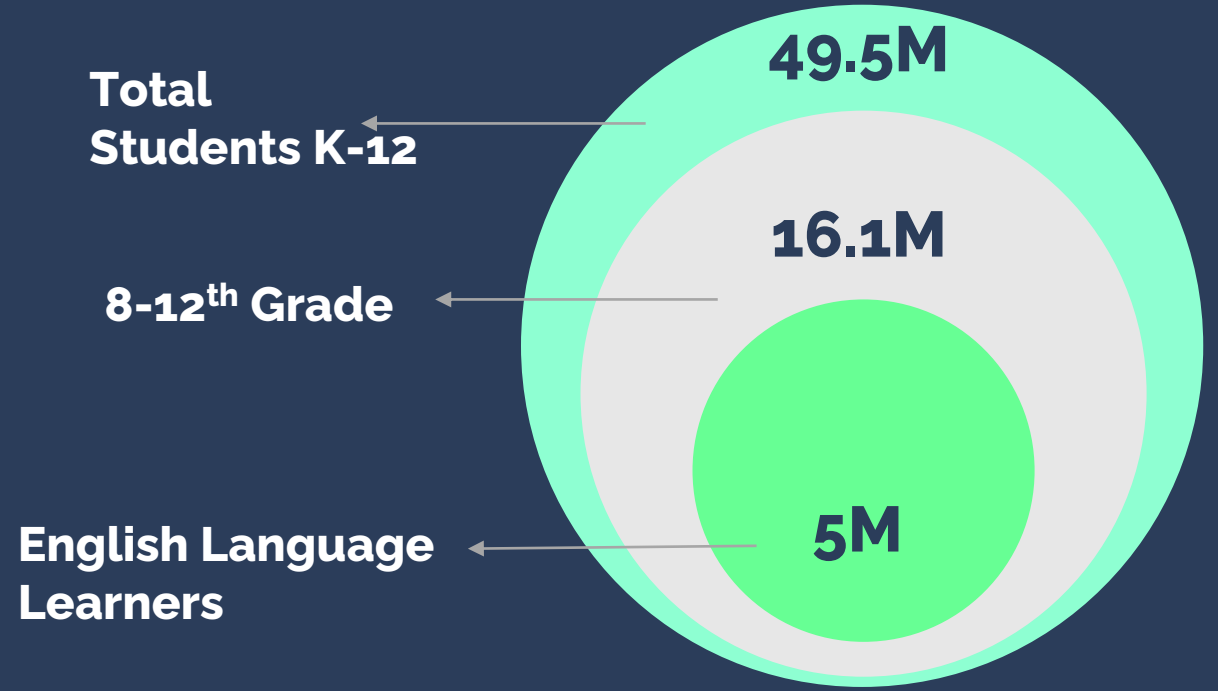
- The objective is to evaluate the English language proficiency of 8th-12th grade students by analyzing their essay dataset.
- To enhance the support for ELLs, a model will be developed using Data Mining techniques
- This model can help reduce the grading burden for teachers and provide better feedback to students on their essays.

Who is an ELL?

- ELLs refer to students who face difficulty communicating effectively and learning in English
- ELLs belong to households and backgrounds where English is not the primary language.
- Due to this, they need special or adjusted instruction to improve their language skills and academic performance.

Market Size

- ELL students are among the fastest growing student population group
- By 2025, 25% of public school students will be ELL students



Data Pre-processing

Data Pre-processing

1 Text Cleaning

a. Apply contractions (don't → do not)

b. Cleaning:
- Lower text, Strip extra spaces, Tabs
- Remove Stop words and Punctuation (and, the, for)

c. Word Tokenization

2 Lemmatize & POS

a. Part of Speech Identification:
- Adverb
- Noun
- Verb
- Adjective

b. Lemmatization (Studying → Study)

3 Spelling Mistakes

a. Identifying Mistakes

b. Correcting Mistakes

4 Count Aggregation

a. Word Count

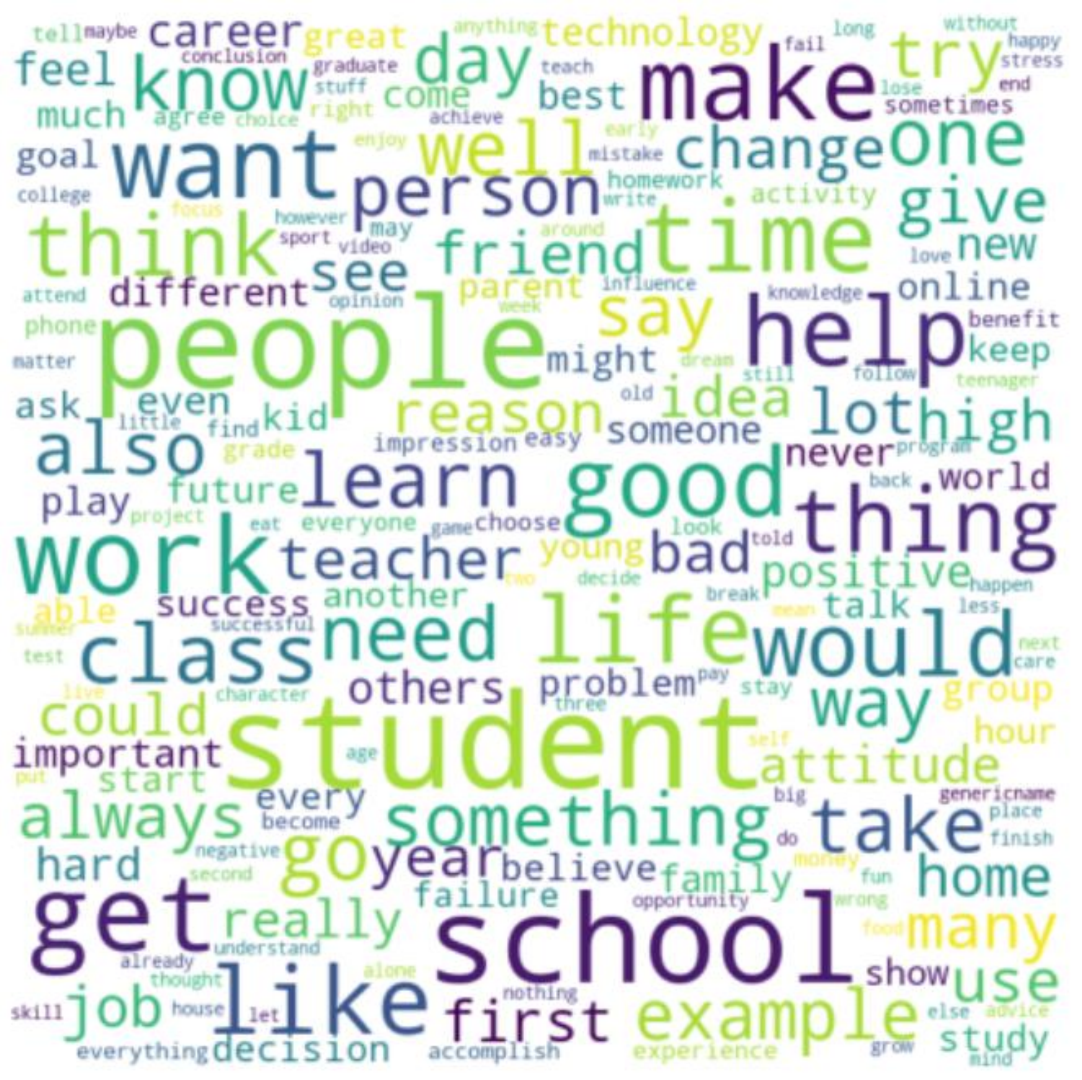
b. Count Spelling Mistakes

c. Count Sentence Metrics
- Length
- Count

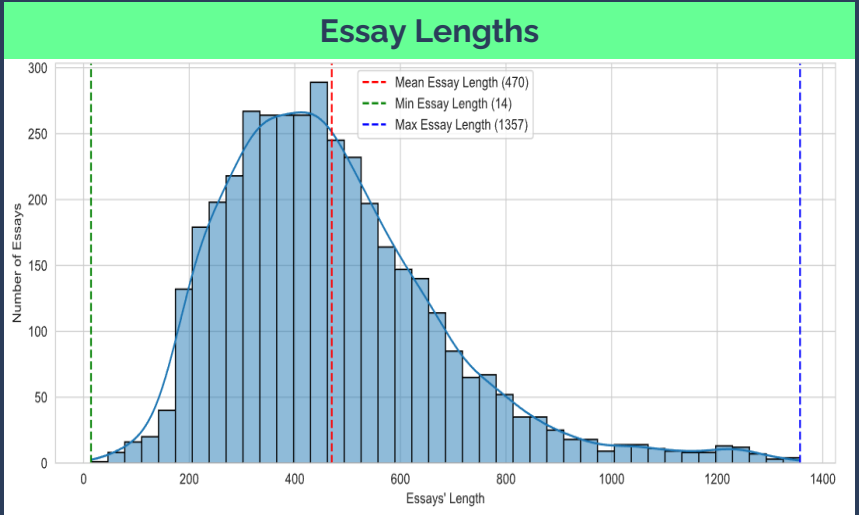
d. POS counts – Noun, Adjective, Adverb & Verb

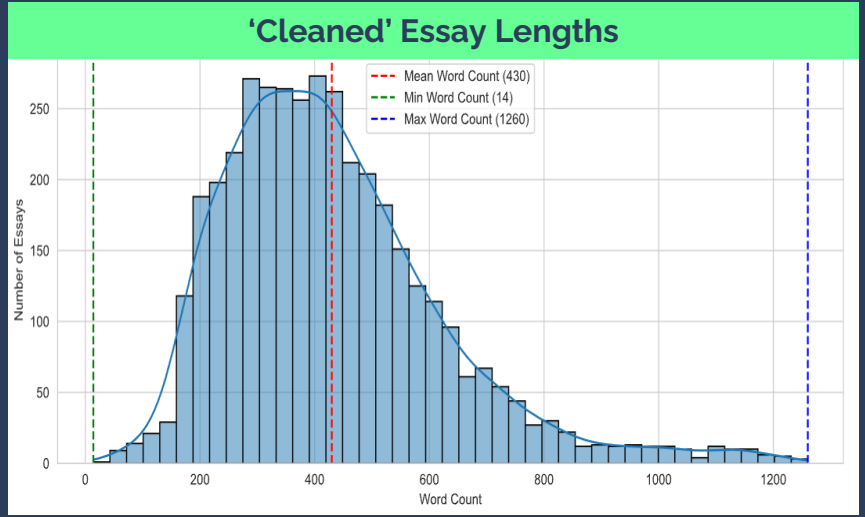
Exploratory Data Analysis

Word Cloud


tell
maybe
career
great
anything
technology
fail
long
without
happy
stress
conclusion
graduate
feel
know
day
teach
make
try
end
stuff
come
best
lose
sometimes
much
agree
choice
right
achieve
want
well
change
one
goal
mistake
person
homework
college
write
activity
give
think
time
however
sport
may
video
friend
new
love
see
different
parent
influence
online
attend
opinion
knowledge
phone
people
say
help
benefit
matter
might
keep
still
follow
ask
even
reason
idea
lot
high
also
kid
find
someone
impression
easy
never
learn
good
future
world
play
back
thing
project
everyone
choose
work
teacher
young
bad
positive
success
another
talk
able
break
happen
less
successful
need
life
would
class
next
care
others
problem
pay
stay
group
could
character
student
way
hour
important
attitude
start
every
something
take
always
become
genericname
place
go
year
believe
family
home
hard
negative
finish
money
fun
failure
get
really
school
many
already
understand
show
use
thought
like
job
first
example
skill
house
let
nothing
study
decision
everything
accomplish
experience
grow

Analyzing Essay Lengths



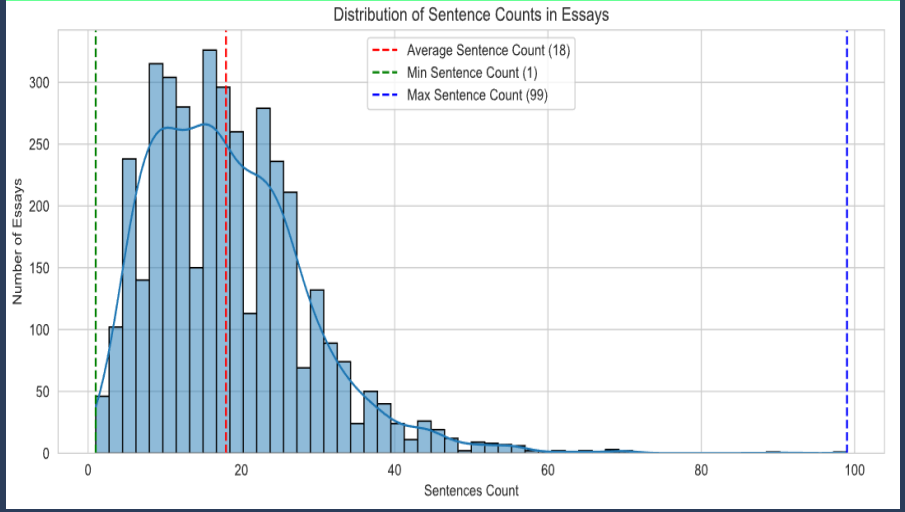


Some students have written very lengthy essays which skews the overall distribution towards right

This visualization is in alignment with the essay length signifying how some students used more words to write essays and made the normal distribution rightly skewed.

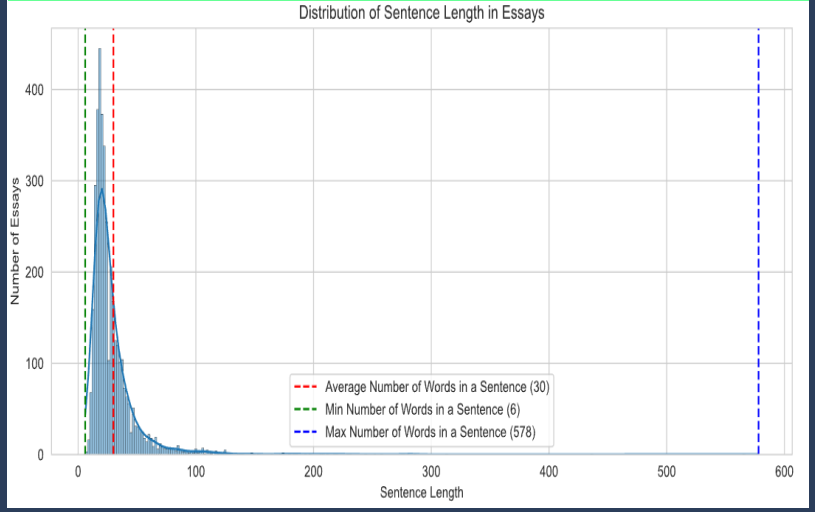
Sentence Analysis

Distribution of Sentence Count



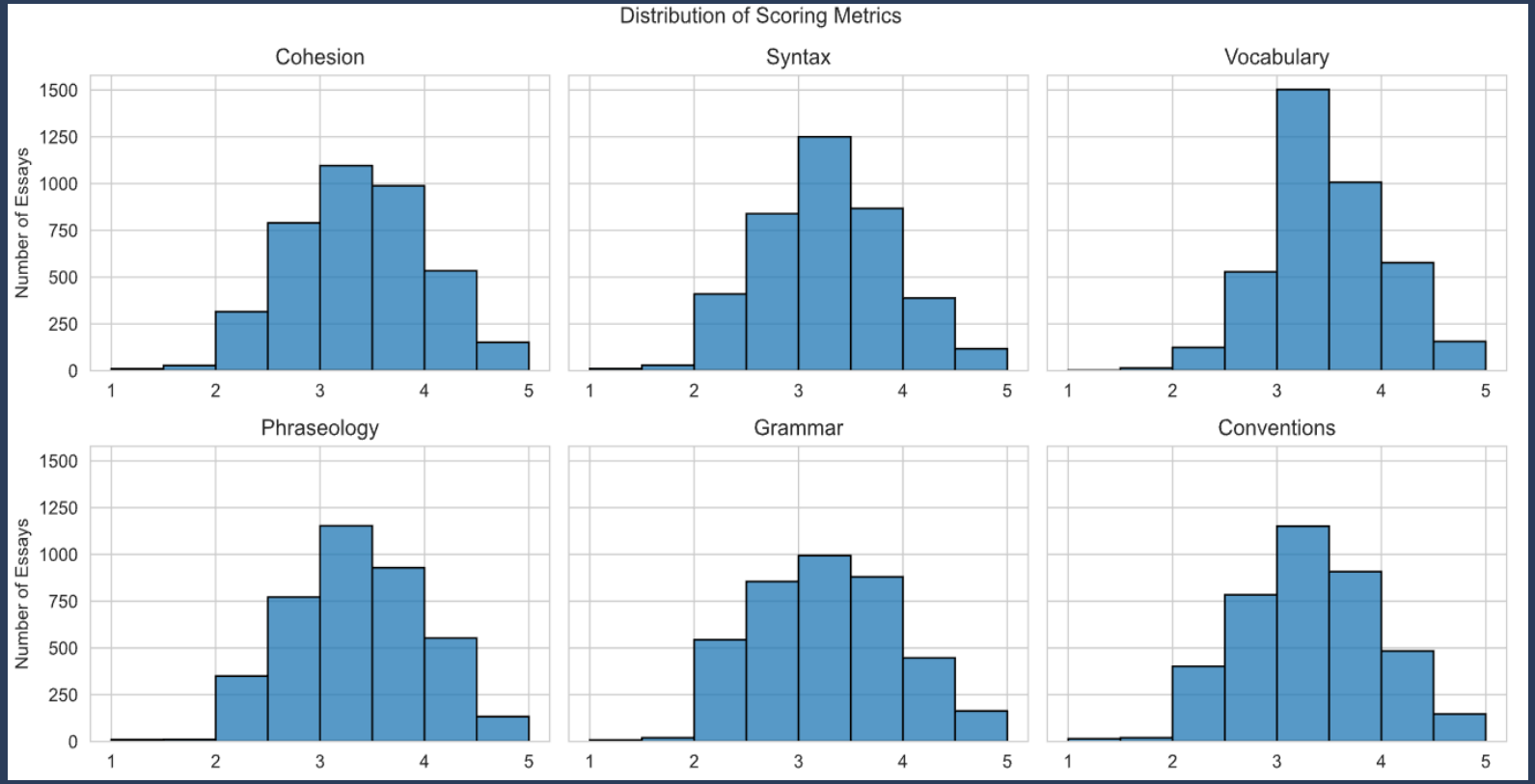
Each student exhibits a distinct writing style, with number of sentences varying in each essay

Distribution of Sentence Length



The distribution of sentences count is highly right-skewed in nature, which signifies how some students wrote exceptionally long sentences.

Essays were graded on 6 metrics

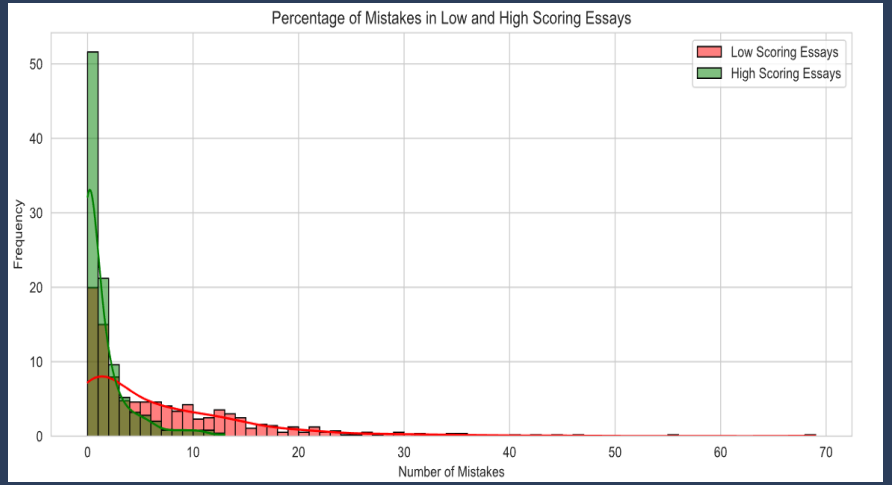


Essays were divided into three scoring groups

ESSAYS

If an essays has the following scores in more than 4 metrics

<= 2.5	>= 4.0	else
Low-Scored Essays	High-Scored Essays	Medium-Scored Essays

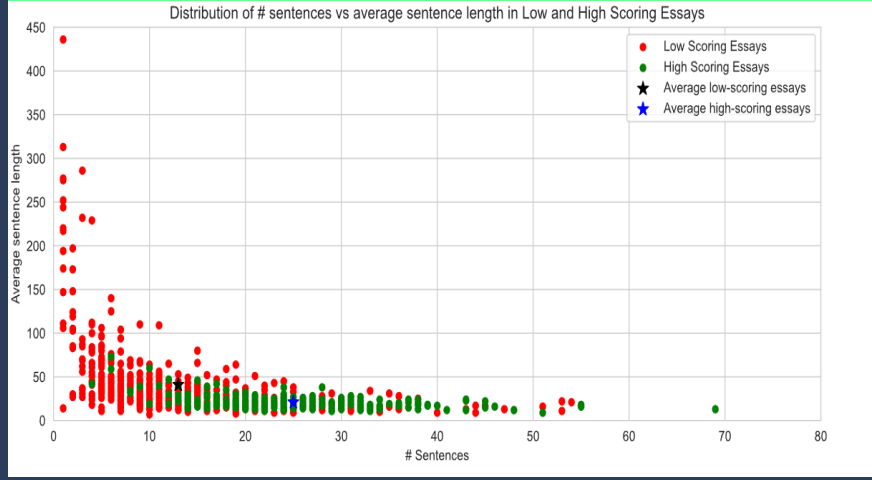


Essays with more spelling mistakes were scored poorly

Different scoring groups

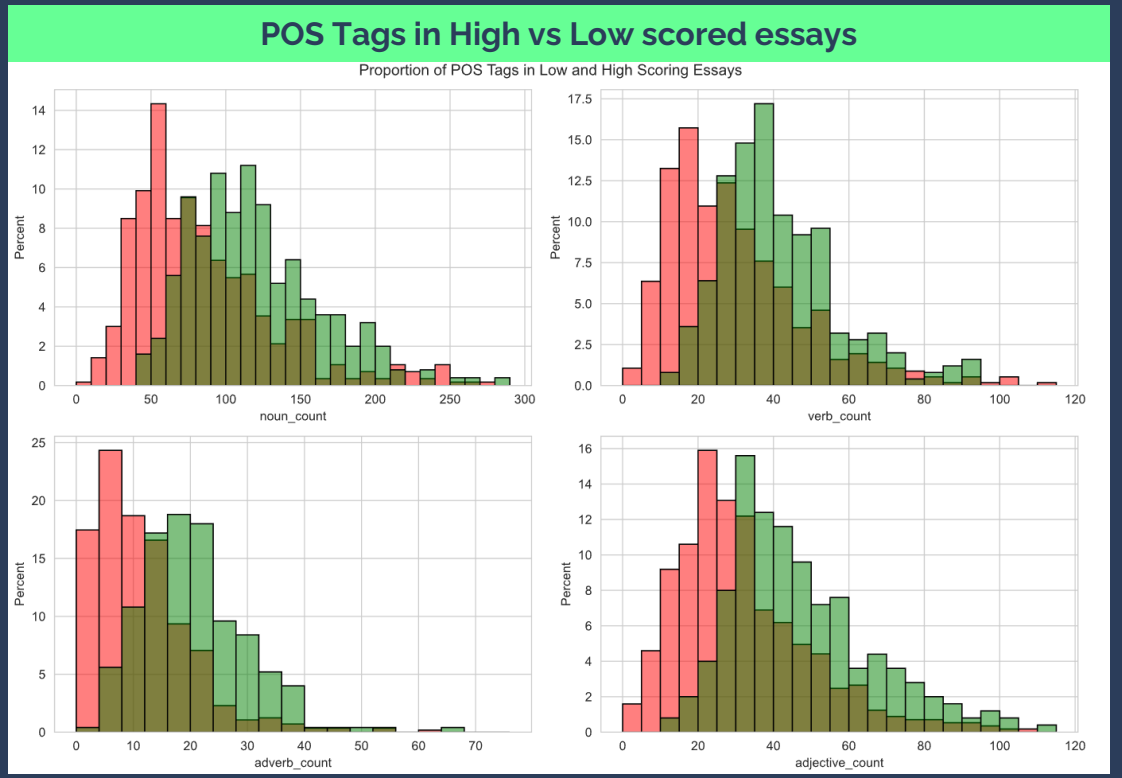
Students with high scoring essays used shorter sentences, whereas students who got low score wrote fewer but lengthy sentences.

Sentence count vs length



Part-of-Speech Tagging

Students who got a high score in their essays, more likely to use more nouns, adverbs, adjectives and verbs.

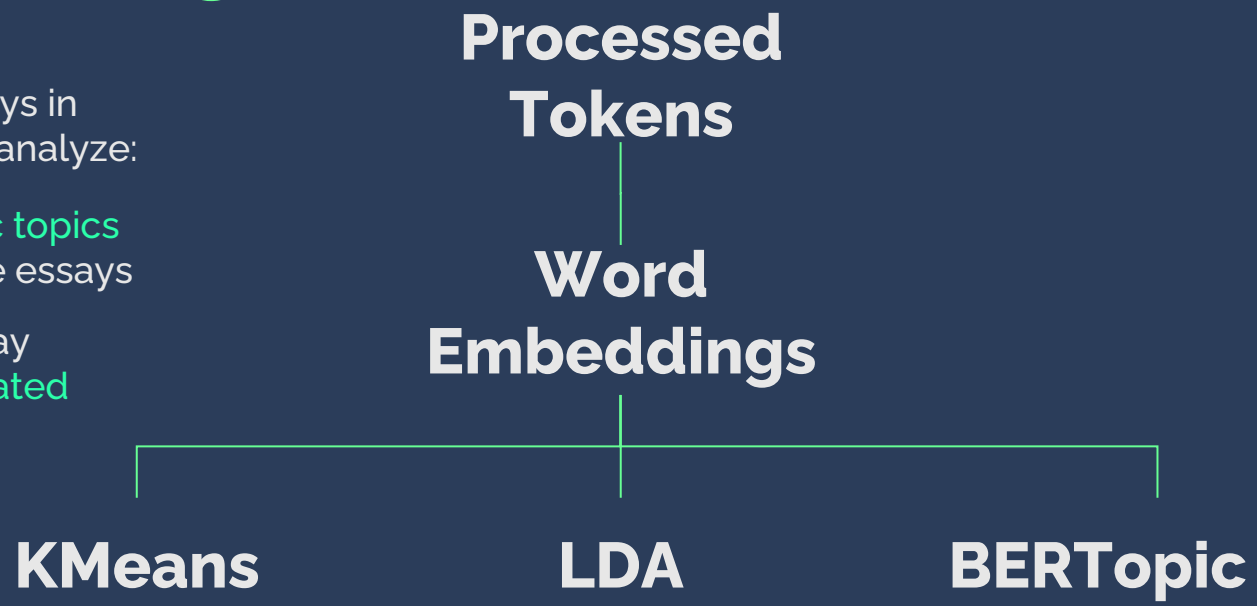


Topic Modeling

Topic Modeling

Having close to 4000 essays in our corpus, we wanted to analyze:

1. If there were specific topics emerging from these essays
2. Whether certain essay topics had an associated scoring bias



Word Embeddings

To convert our list of lemmatized text into numerical representation, we implemented 5 embedding techniques.

BOW

Creates an unordered set (Bag) of words – only accounts for word presence or absence

Count Vectorizer

Uses BOW to create a matrix using the frequency of each word in the document

TF-IDF

Evaluates the importance of a word in a document by multiplying the frequency (TF) with the logarithm of the inverse frequency of that word in the corpus (IDF)

Word2Vec

Trains a neural network to predict the probability of a word based on its context or the probability of a context given a word.

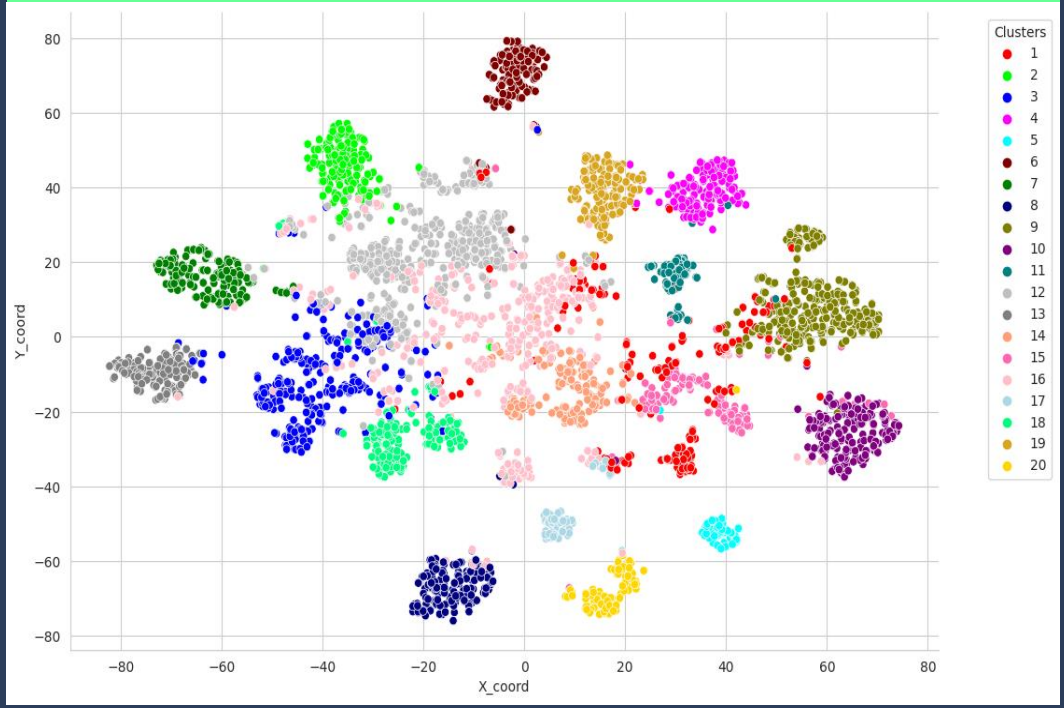
BERT

Pre-trained large language model (developed by Google) trained on a diverse range of text data, including Wikipedia articles, books, and web pages

Word2Vec with KMeans

- Created Word2Vec embeddings for each tokenized word (corrected) in our corpus
- Translated word-level embeddings to document-level embeddings by averaging the word2vec embeddings for each word
- Implemented KMeans clustering for n=20 clusters

KMeans visualization with TSNE



Few emerging Topics

(with most representative keywords)

Cluster 11
Vacation and Summer Breaks

summer vacation break winter longer evening period long spring shorter

Cluster 05
Food and Meals

cafeteria eat healthier food lunch menu meal starve healthy vegetable

Cluster 08
Impact of Technology

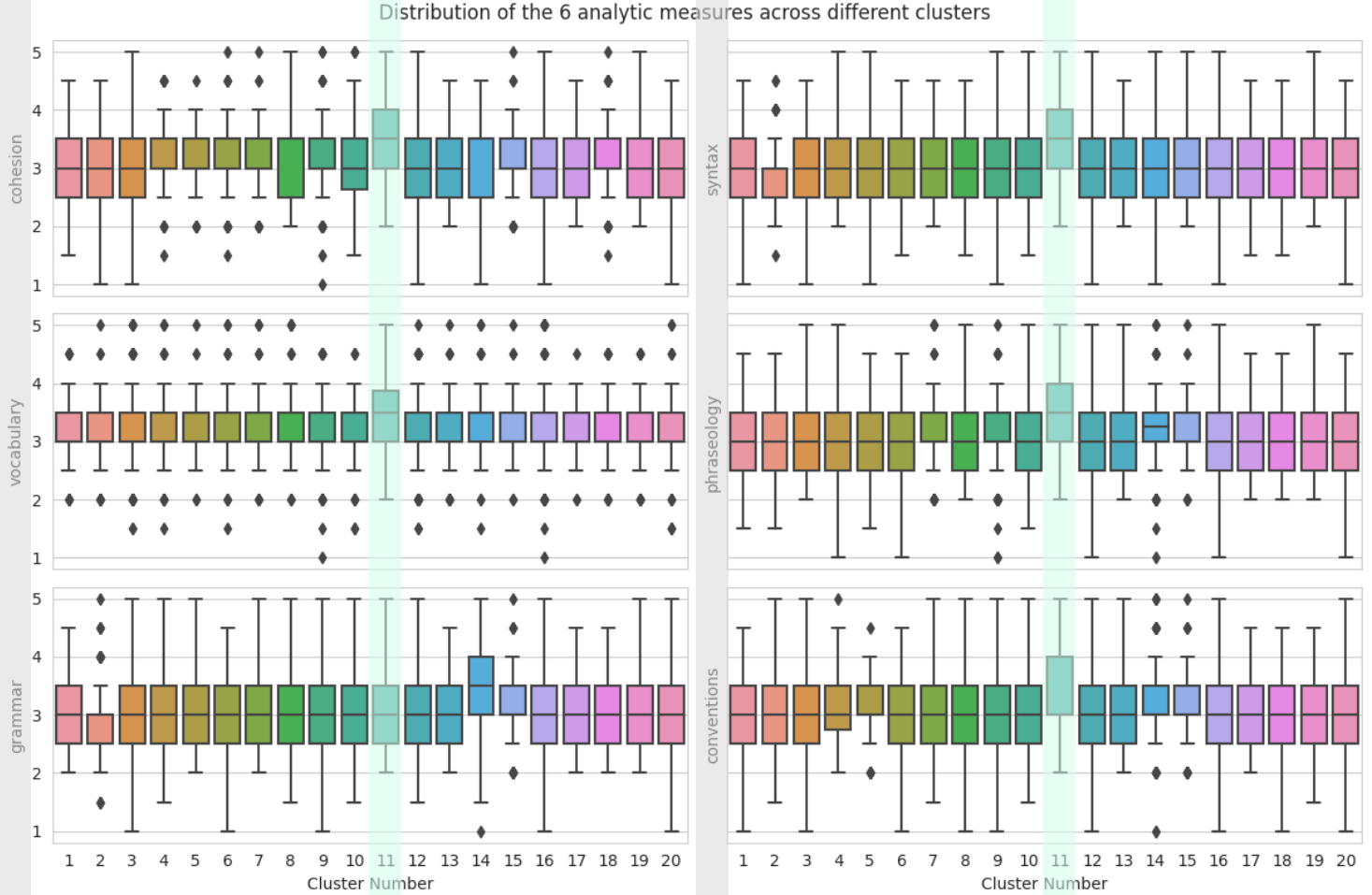
technology contact device social medium today limitation addict human tool

Cluster 14
Hobbies and Vocational activities

swim sing bike dance lionel football basketball ride ball guitar

We do not observe any scoring bias in any of the topic clusters

Cluster 11 (vacation) does show slightly higher scores



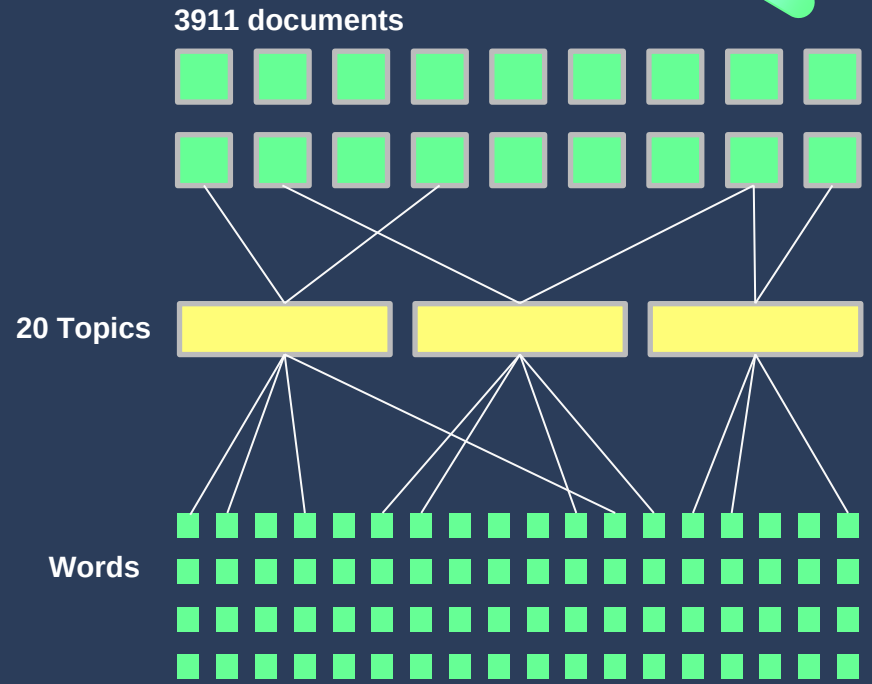
Topic Modeling - LDA

We tried to identify 20 different topics using LDA. LDA algorithm finds the weight of connections between documents and topics and between topics and words.

It started off by randomly assigning topics to each word in our essays. It then calculates a document to topic count.

The model converges by re-assigning the topic of each word, so that each word is assigned to min number of topics

Each topic is a list of probabilities of words
Each document is a list of probabilities of topics



Few LDA Topics

(with most representative keywords)

Topic 14
Goal Setting

goal, could, achieve, high, set, may, would, great, aim, low

Topic 12
Self-Esteem

work, student, praise, achievement, believe, skill, improve, self-esteem, confidence

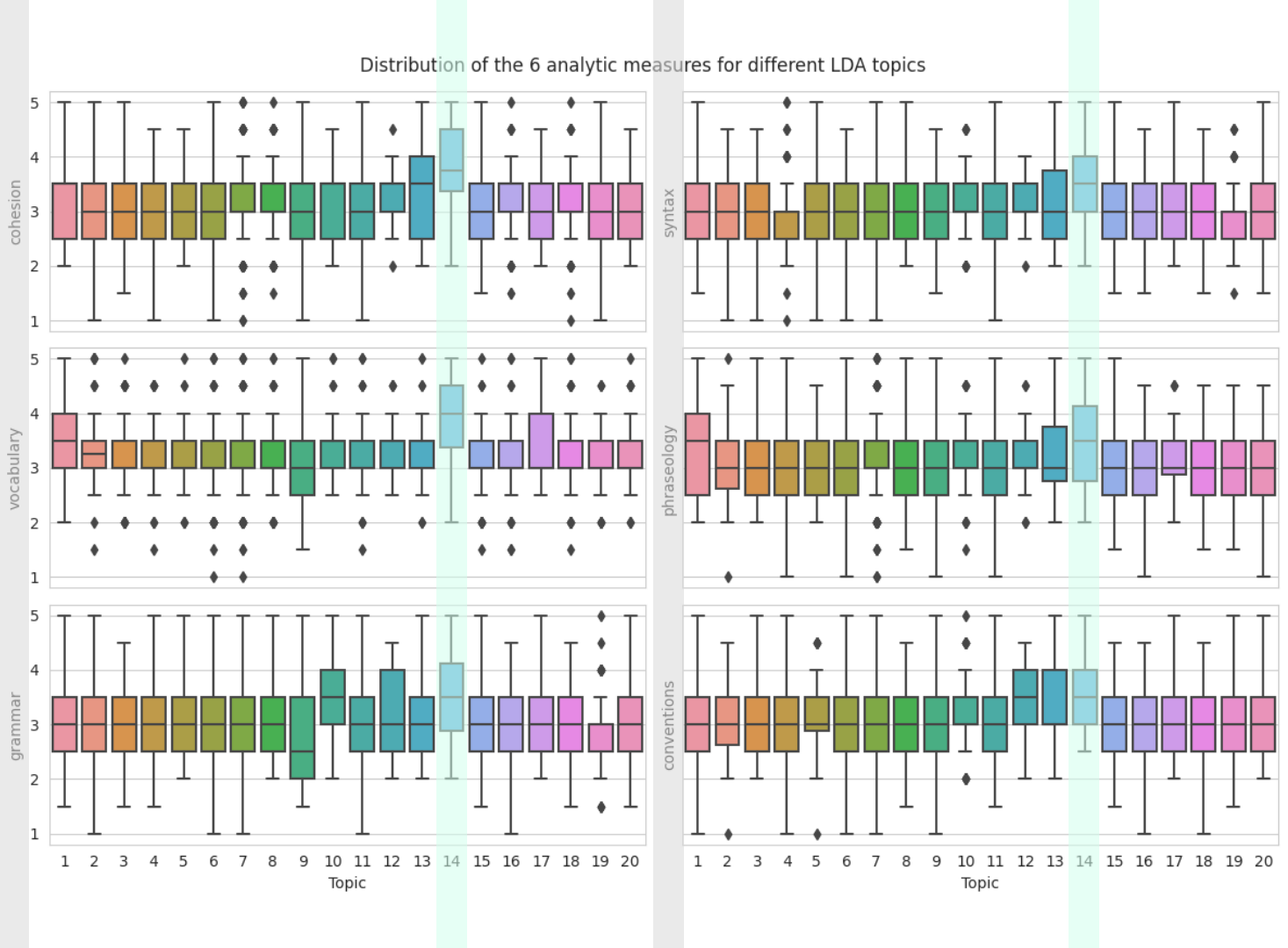
Topic 06
Importance of trying new things

try, thing, get, make, something, never, know, learn, new, life

Topic 17
Imagination and Creativity

knowledge, imagination, art, use, imagine, important, think, draw, create, make

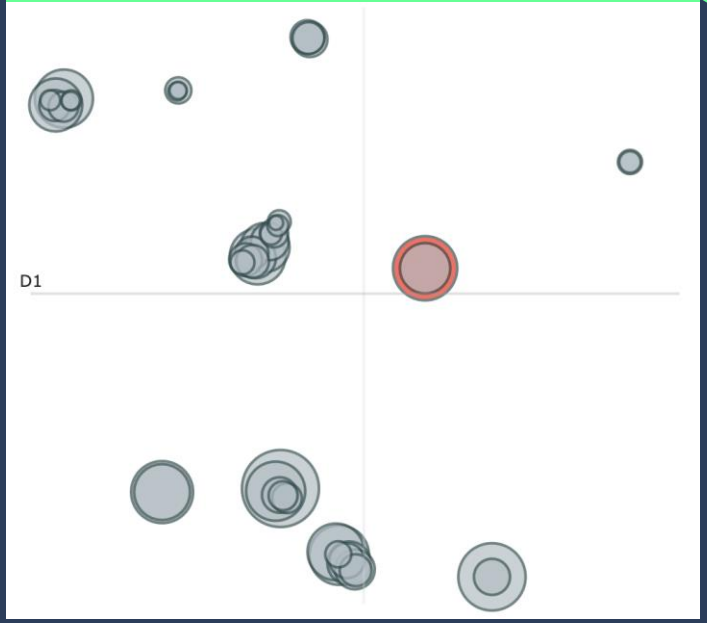
Topic 14 (Goal setting) did show consistently higher scores in all analytic measures



BERT (Experiment)

- BERT takes into account the entire context of a word in a sentence (document), rather than just the words that come before or after it
- BERT Topic applies a clustering algorithm like HDBSCAN to BERT embedded (and dimensionality reduced) data to group similar documents into topics
- It can automatically determine the optimal number of topics without manual tweaking

BERT identified 47 topics
(Intertopic distances below)



Interesting BERT Topics

(with most representative keywords)

Topic 0
Online learning

online, class, home, attend, student, school, benefit, take, learn, video

Topic 15
Homework and Assignments

homework, club, teacher, help, student, school, afterschool, get, grade, understand

Topic 29
Community Service

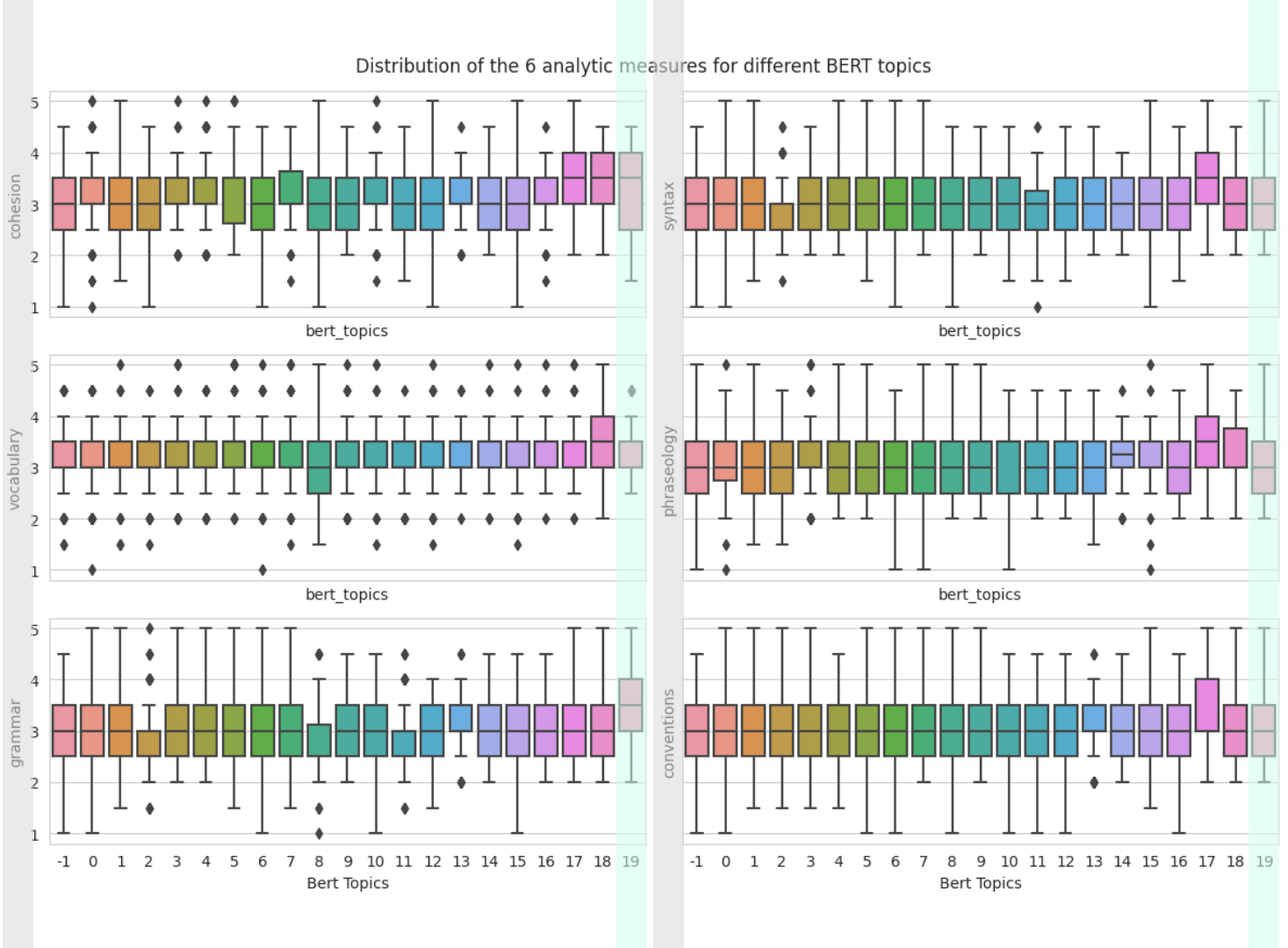
community, service, clean, help, principal, dear, perform, trash, litter

Topic 40
Job Market

job, hire, employee, employer, work, experience, hard, responsible, company, customer

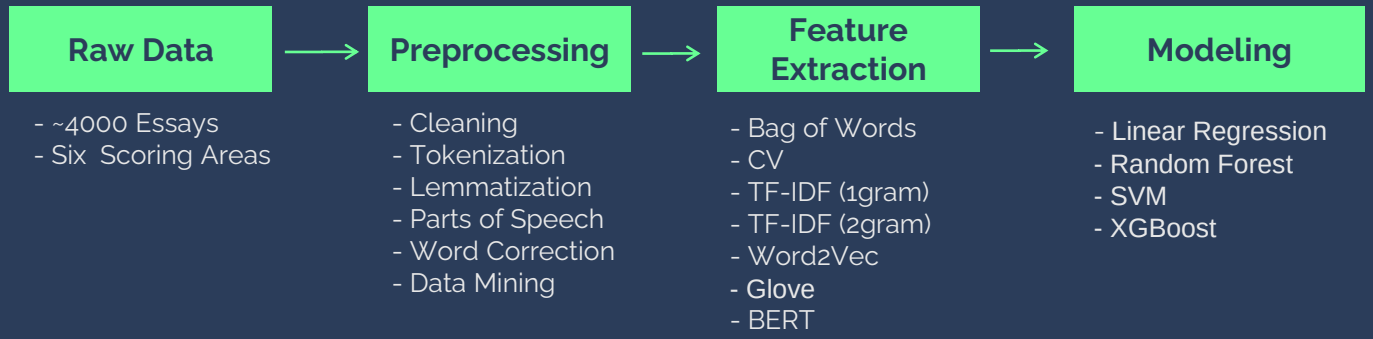
Scores for top 20 Bert topics

No topic was showing a particular scoring bias

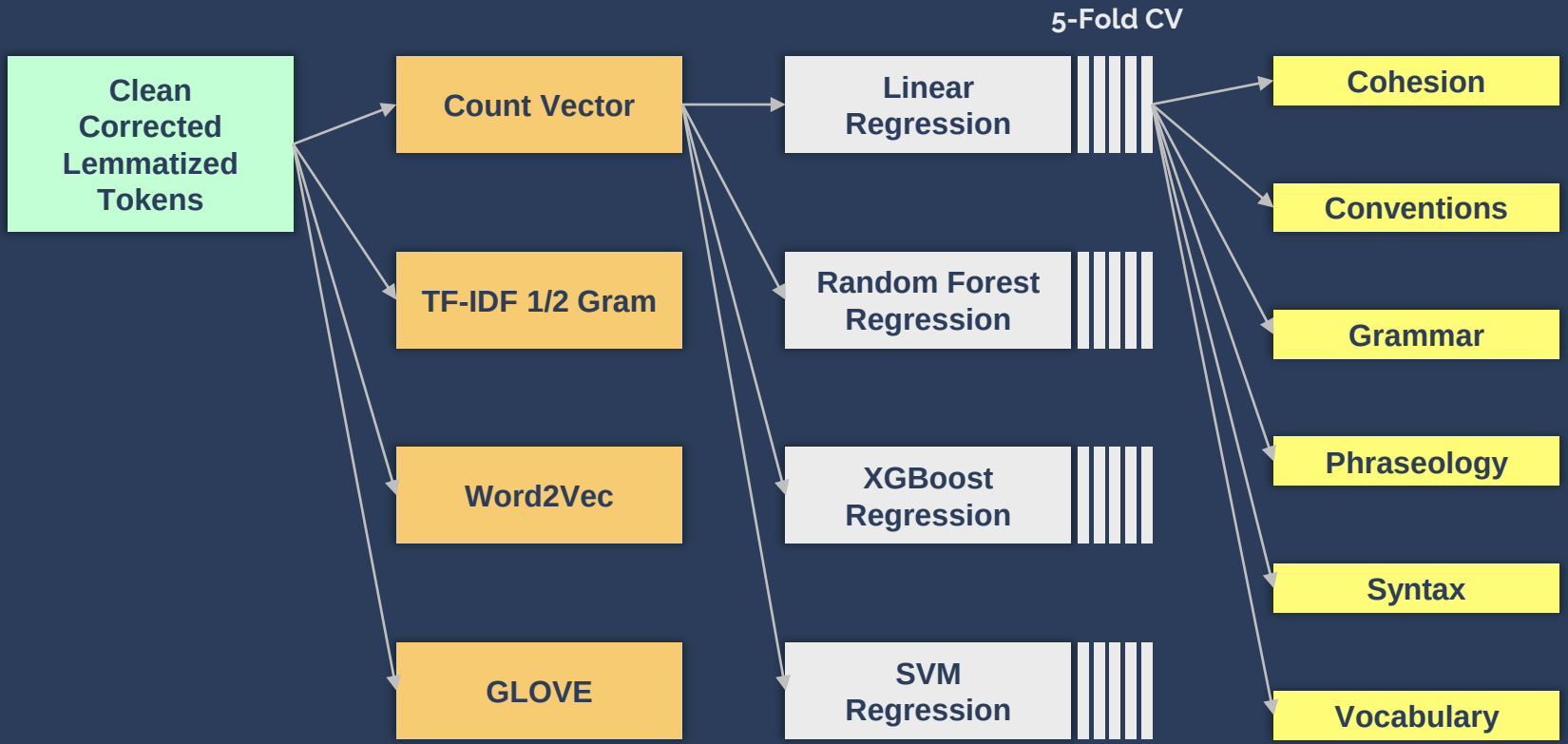


Modeling

Data Mining Pipeline

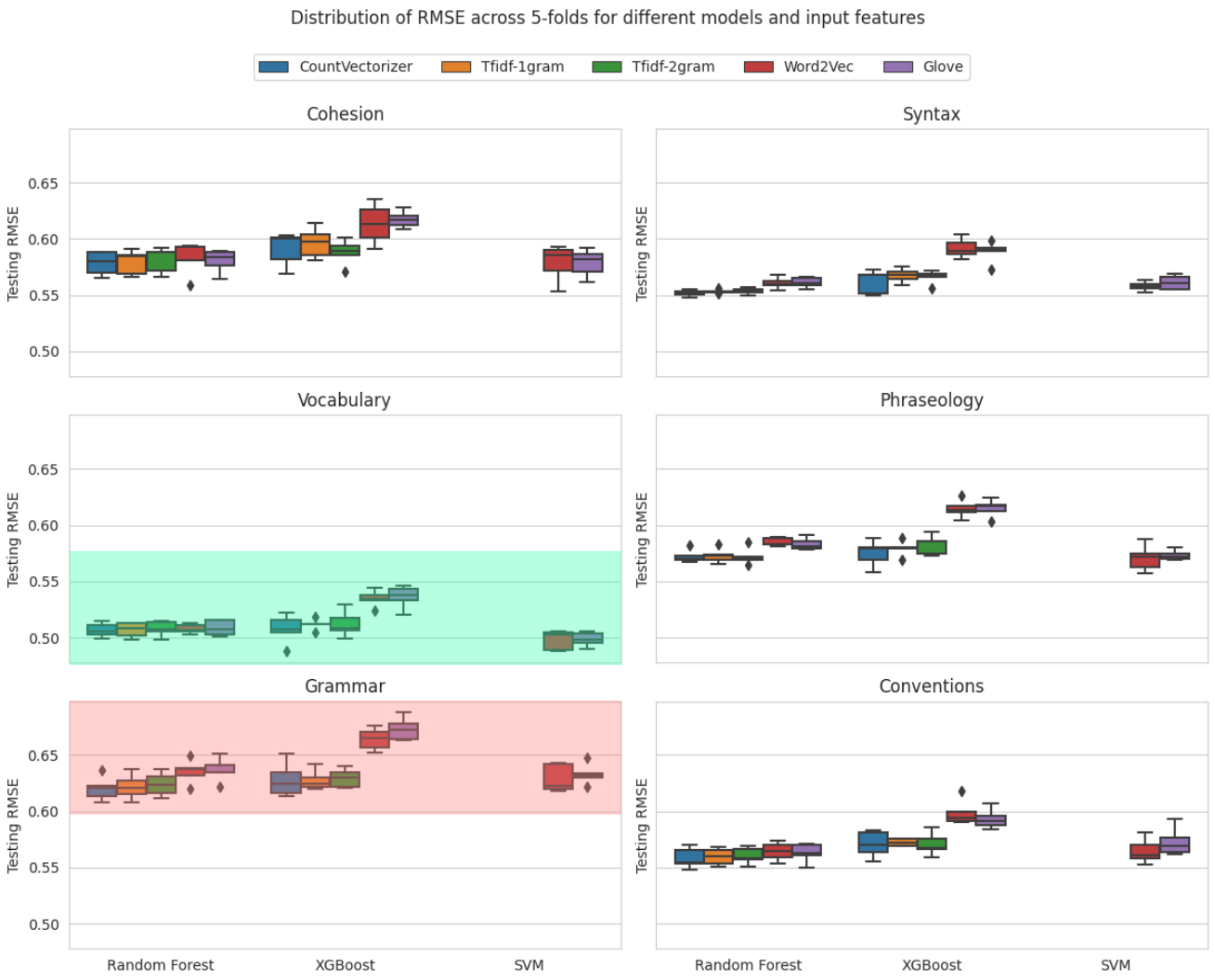


Modeling Approach



Modeling Results

Some measures are more difficult to model than the others



Modeling Results

Target	Embedding	Model	RMSE	MAE
Cohesion	CountVectorizer	Random Forest	0.58	0.47
Conventions	CountVectorizer	Random Forest	0.56	0.44
Grammar	CountVectorizer	Random Forest	0.62	0.50
Phraseology	Word2Vec	SVM	0.57	0.45
Syntax	CountVectorizer	Random Forest	0.55	0.44
Vocabulary	Glove	SVM	0.49	0.39

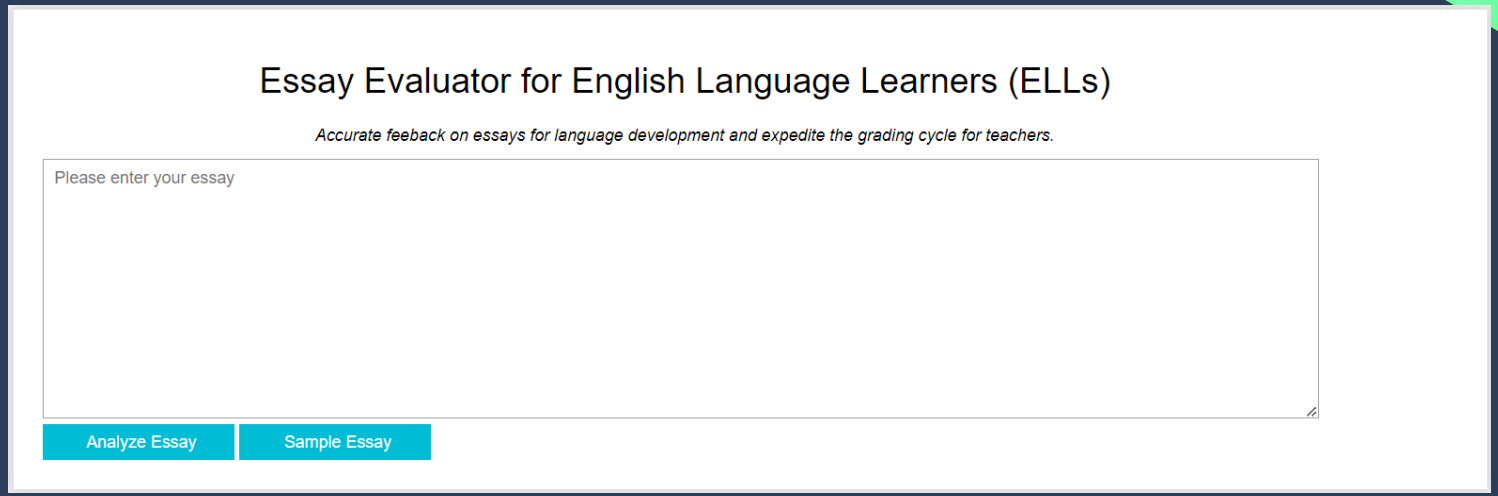
Results and Recommendations

Results

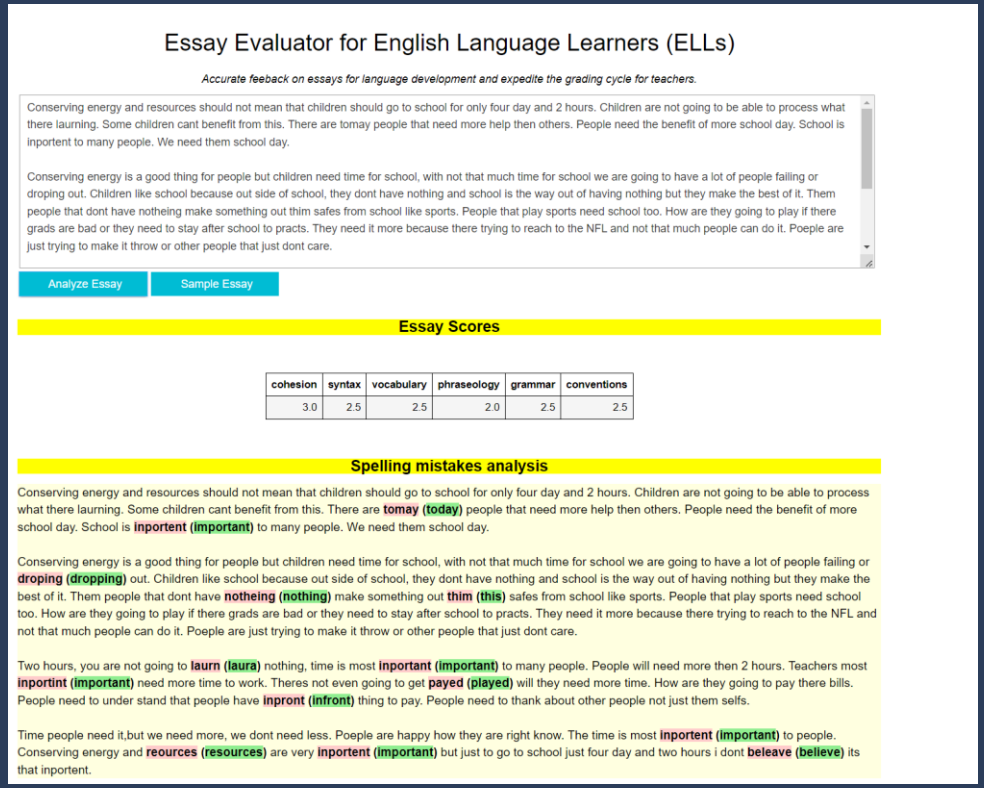
We built an interactive app

- This tool can be used by graders and students both to analyze their essay submissions
- The tool is powered by the Random Forest Model we obtained from our training data
- Given this tool is built for ELLs, it provides many learning features that can help them improve their writing

App UI


Essay Evaluator for English Language Learners (ELLs)
Accurate feeback on essays for language development and expedite the grading cycle for teachers.
Please enter your essay
Analyze Essay
Sample Essay

App UI (contd.)



Essay Evaluator for English Language Learners (ELLs)

Accurate feeback on essays for language development and expedite the grading cycle for teachers.

Conserving energy and resources should not mean that children should go to school for only four day and 2 hours. Children are not going to be able to process what there laurning. Some children cant benefit from this. There are tomay people that need more help then others. People need the benefit of more school day. School is inportent to many people. We need them school day.

Conserving energy is a good thing for people but children need time for school, with not that much time for school we are going to have a lot of people failing or droping out. Children like school because out side of school, they dont have nothing and school is the way out of having nothing but they make the best of it. Them people that dont have notheing make something out thim safes from school like sports. People that play sports need school too. How are they going to play if there grads are bad or they need to stay after school to practs. They need it more because there trying to reach to the NFL and not that much people can do it. Poeple are just trying to make it throw or other people that just dont care.

Analyze Essay | Sample Essay

Essay Scores

cohesion	syntax	vocabulary	phraseology	grammar	conventions
3.0	2.5	2.5	2.0	2.5	2.5

Spelling mistakes analysis

Conserving energy and resources should not mean that children should go to school for only four day and 2 hours. Children are not going to be able to process what there laurning. Some children cant benefit from this. There are **tomay** (**today**) people that need more help then others. People need the benefit of more school day. School is **inportent** (**important**) to many people. We need them school day.

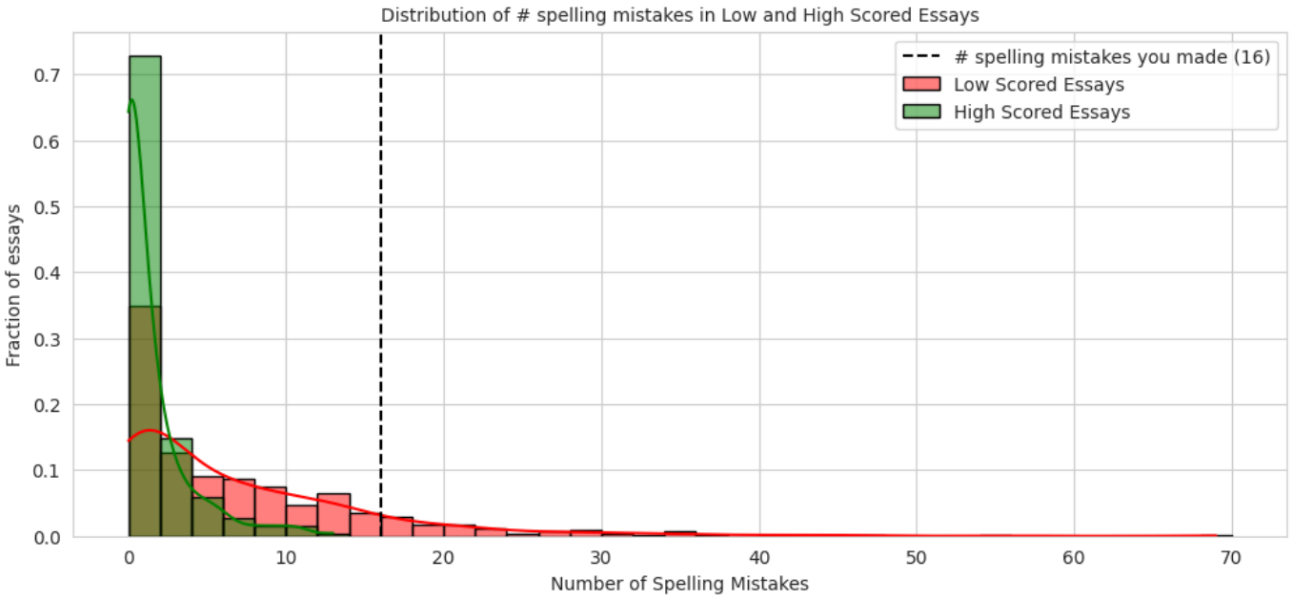
Conserving energy is a good thing for people but children need time for school, with not that much time for school we are going to have a lot of people failing or **droping** (**dropping**) out. Children like school because out side of school, they dont have nothing and school is the way out of having nothing but they make the best of it. Them people that dont have **notheing** (**nothing**) make something out **thim** (**this**) safes from school like sports. People that play sports need school too. How are they going to play if there grads are bad or they need to stay after school to practs. They need it more because there trying to reach to the NFL and not that much people can do it. Poeple are just trying to make it throw or other people that just dont care.

Two hours, you are not going to **laurn** (**laura**) nothing, time is most **inportant** (**important**) to many people. People will need more then 2 hours. Teachers most **inportint** (**important**) need more time to work. Theres not even going to get **payed** (**played**) will they need more time. How are they going to pay there bills. People need to under stand that people have **inpront** (**infront**) thing to pay. People need to thank about other people not just them selfs.

Time people need it,but we need more, we dont need less. Poeple are happy how they are right know. The time is most **inportent** (**important**) to people. Conserving energy and **reources** (**resources**) are very **inportent** (**important**) but just to go to school just four day and two hours i dont **beleave** (**believe**) its that inportent.

Recommendations

Essays with more than 10 spelling mistakes tend to score very low. You've made 16 spelling mistakes in your essay.



Thank you