

Cross-Lingual Embeddings Analysis Across Sentence Complexity and Language Groups

Fall 2024

Akshat Shah, Siddhi Bansal



Research Question

How does **sentence complexity** affect the similarity between **cross-lingual embeddings** across different **embedding models**?

Sentence complexity

Languages

Vectorizers

Motivation + Purpose



Being multilingual ourselves, we were curious to explore how embeddings function across different languages



Analyze the **efficacy of cross-lingual embedding models** on different types of sentences



Investigate which **languages exhibit the greatest similarity** based on cosine similarities in sentence complexity.

Background

VECTORIZERS



OpenAI



LASER



LaBSE

SENTENCE TYPES



Simple



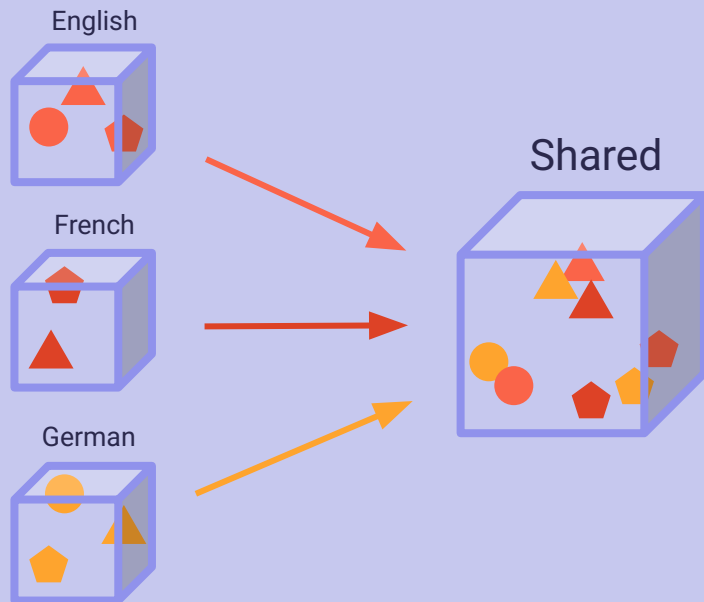
Compound



Complex

Background

Cross-Lingual Embeddings



Purpose

Represent sentences across languages in the same vector space

Applications

Translation, semantic search, and crosslingual understanding

Methods



Data Collection

Languages: 60 languages (as supported by our vectorizers)

Sentences: 300 total (100 each: simple, complex, compound) sourced from educational websites teaching sentence types



























Analysis Process

1. Translate all sentences into 60 languages with Google Translate API
2. Generate embeddings for translated sentences
3. Compute the pairwise cosine similarity of sentence embeddings for different languages, grouped by sentence complexity
4. Visualize the most similar languages

Most Similar Languages

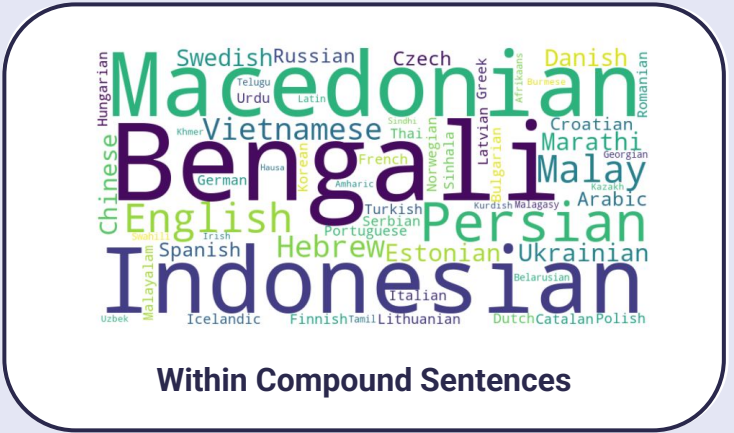
binned by sentence complexity

LANGUAGE	SIMPLE	COMPOUND	COMPLEX
English	  	  	  
German	  	  	  
Spanish	  	  	  

 spanish  portuguese  german  danish  dutch  swedish  english
 romanian  bulgarian



Within Complex Sentences



Across All Sentences

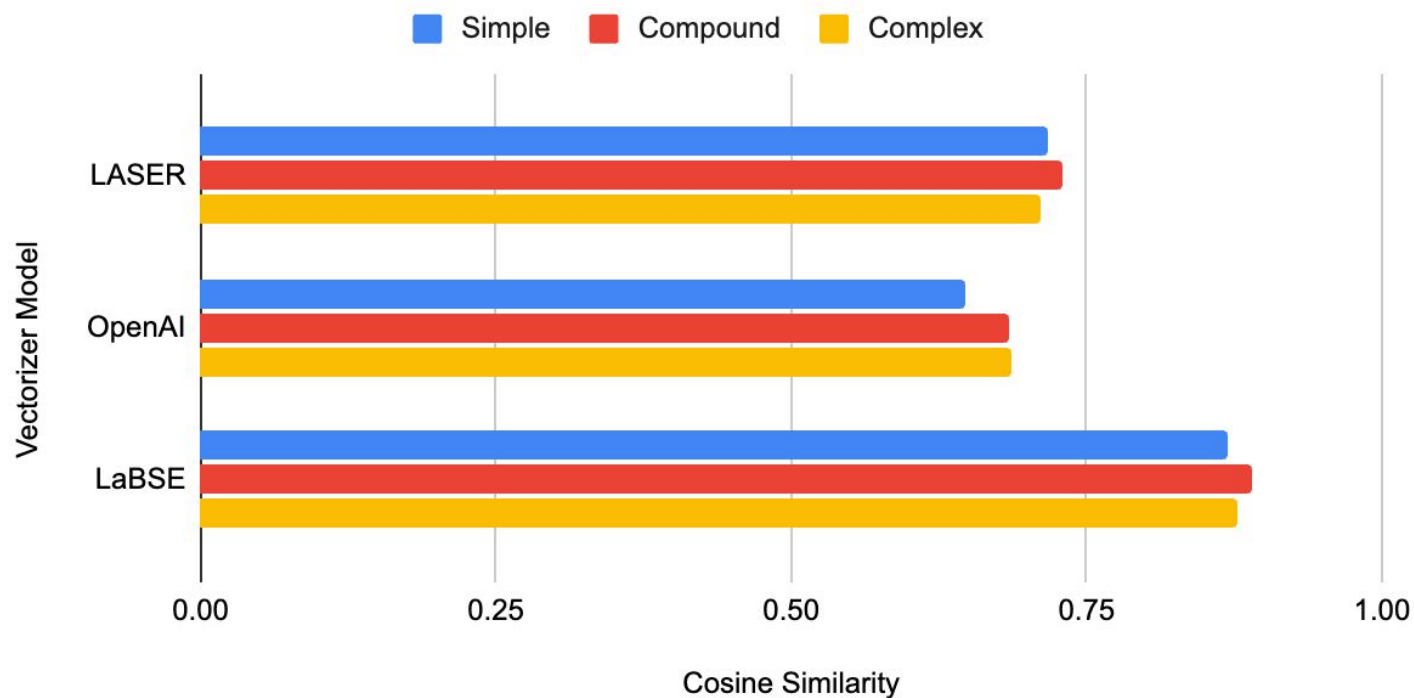


Within Complex Sentences



Across All Sentences

Average Cosine Similarities of Sentences by Model and Sentence Type



Findings

Comparing Models

OpenAI's embeddings yielded lower cosine similarities compared to LASER and LaBSE.

Sentence Complexity & Language Similarity

Our findings indicate that sentence complexity does not impact cosine similarity when examining the similarity between cross-lingual embeddings across different embedding models.

Most Universally Similar Languages

1. Macedonian



2. Bulgarian 

3. Danish 

Limitations

Translations

Translation accuracy can vary, leading to discrepancies across languages.

Sentences Corpus

Manually collecting data limits the dataset size and introduces some bias.

Cross-Lingual Vectorizer

The vectorizers may struggle with nuanced language differences, also limited to certain languages.

Confounding Variables

Uncontrolled confounding factors can distort the results.

Extensions

Corpus-Specific Testing

Test models on specific corpora (e.g., Bible passages) to evaluate performance on unique text types.

Fine-Tuning Vectorizers

Adjust vectorization methods for better embeddings tailored to specific text features.

Embedding Effectiveness vs Language Popularity

Study how embedding quality varies with the popularity of a language.

Thank You!

¡Gracias!



in Spanish

Danke!



in German

धन्यवाद!



in Hindi

Merci!

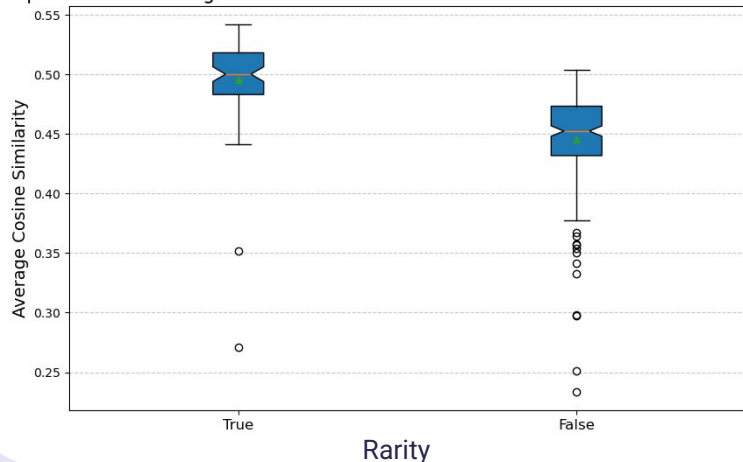


in French

Consideration: Confounding Variable

Confounding Variable: Frequent words are more robustly encoded than infrequent ones

Spread of Cross-Lingual Cosine Similarities for Sentences Across Sentence Rarity



Term Frequency

wordfrequency module: identify frequent and infrequent words

rarity threshold based on average sentence word frequency → classify sentence as "rare" or "not rare"

WHOA!

This could be the part of the presentation where you can introduce yourself, write your email...





Ni hao!

**“THIS IS A QUOTE, WORDS
FULL OF WISDOM THAT
SOMEONE IMPORTANT SAID
AND CAN MAKE THE
READER GET INSPIRED.”**

—Someone Famous

TABLE OF CONTENTS

01

FIRST SECTION

You could describe
the topic of the
presentation here

02

SECOND SECTION

You could describe
the topic of the
presentation here

03

THIRD SECTION

You could describe
the topic of the
presentation here

INTRODUCTION

Mercury is the closest planet to the Sun and the smallest one in the Solar System—it's only a bit larger than the Moon

Hi!





¡Hola!

01

FIRST SECTION

You can enter a subtitle
here if you need it

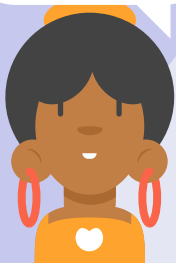
THE SLIDE TITLE GOES HERE!

Do you know what helps you make your point clear?
Lists like this one:

- They're simple
- You can organize your ideas clearly
- You'll never forget to buy milk!

The audience won't miss the point of your presentation

Ni hao!



Kia Ora!



MAYBE YOU NEED TO DIVIDE THE CONTENT



MERCURY

Mercury is the closest planet to the Sun and the smallest one



VENUS

Venus has a beautiful name and is the second planet from the Sun

YOU COULD USE THREE COLUMNS



MARS

Despite being red,
Mars is a cold place



JUPITER

Jupiter is a gas giant
and the biggest



SATURN

Saturn is a gas giant
and has rings



Hi!

A PICTURE REINFORCES THE CONCEPT

Images reveal large amounts of data, so remember: use an image instead of a long text



**A PICTURE IS WORTH A
THOUSAND WORDS!**

AWESOME WORDS



MOST SPOKEN MOTHER LANGUAGES



To modify this graph, click on it, follow the link, change the data and paste the resulting graph here, replacing this one

ABOUT THE TONGUES

LANGUAGE	FAMILY	SPEAKERS
English	Indo-European	1.268 BI
Mandarin	Sino-Tibetan	1.120 BI
Hindi	Indo-European	637.3 MI
Spanish	Indo-European	537.9 MI

PERCENTAGE OF NATIVE SPEAKERS

35%



Venus has a
beautiful name

65%



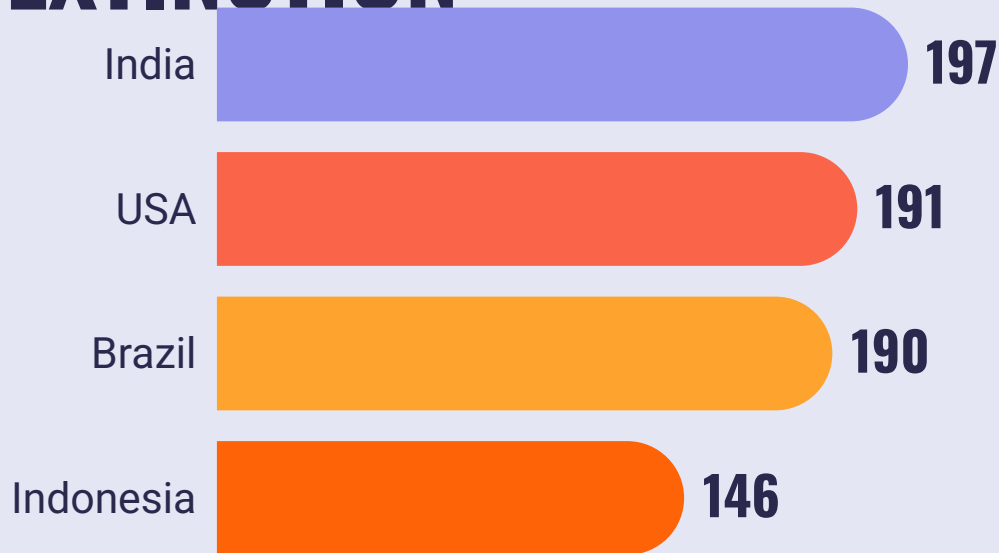
Jupiter is the
biggest planet

85%



Mars is a very
cold place

COUNTRIES WITH LANGUAGES AT RISK OF EXTINCTION



101

countries speak
English all over
the world

02

SECOND SECTION

You can enter a subtitle
here if you need it

Jambo!

A stylized illustration of a person with dark skin and large, voluminous orange curly hair. They are wearing a white tank top under a dark blue long-sleeved shirt. A white speech bubble with a dark blue outline is positioned above their head, containing the text 'Jambo!'. The background is a light purple color with several white, horizontal, brushstroke-like lines scattered around the person's head and shoulders.

ABOUT AFRICAN LANGUAGES

MERCURY

Mercury is the closest planet to the Sun

MARS

Despite being red, Mars is a cold place

JUPITER

It's the biggest planet in the Solar System

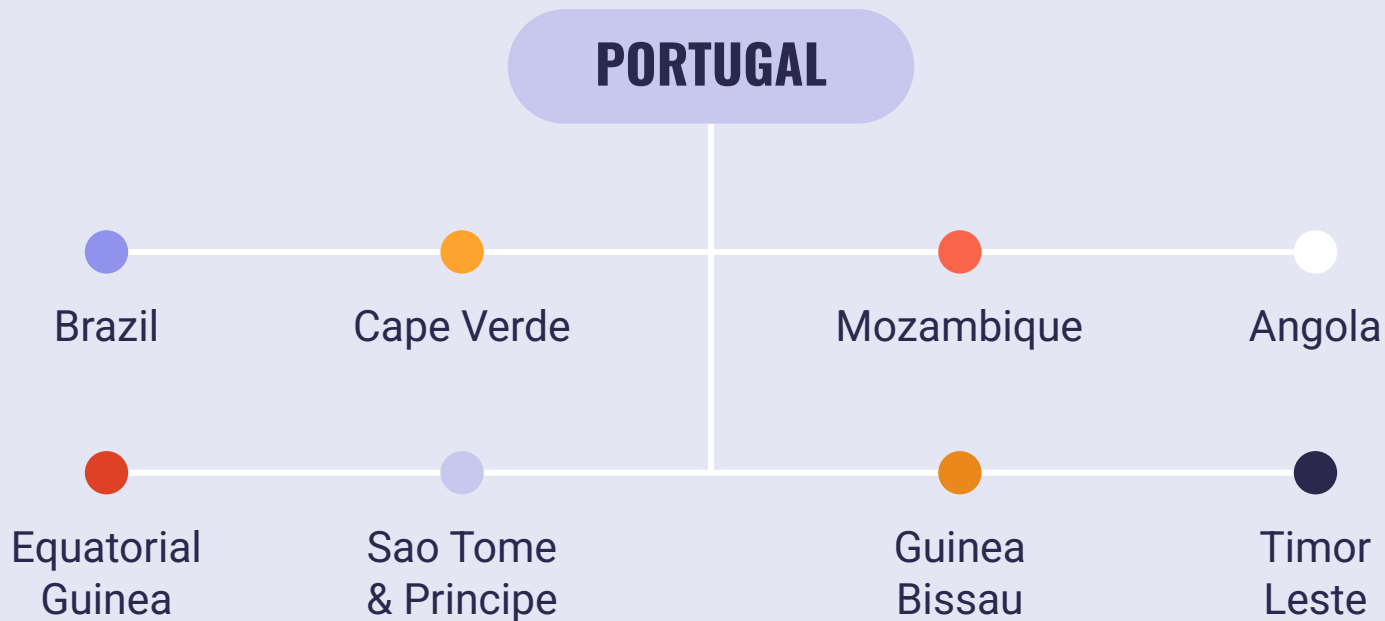
SATURN

Saturn is a gas giant and has rings

NEPTUNE

It's the farthest planet from the Sun

PORTUGUESE SPEAKING COUNTRIES



ASPECTS OF MOTHER TONGUES

01

POSITIVE ASPECTS

Venus has a beautiful name, but it's terribly hot

Despite being red, Mars is actually a cold place

02

NEGATIVE ASPECTS

Saturn is the ringed one and a gas giant

Neptune is the farthest planet from the Sun

MORE TEXTS ABOUT THIS

SATURN

It's a gas giant
and has rings

NEPTUNE

Neptune is the
farthest planet

MERCURY

Mercury is the
closest planet

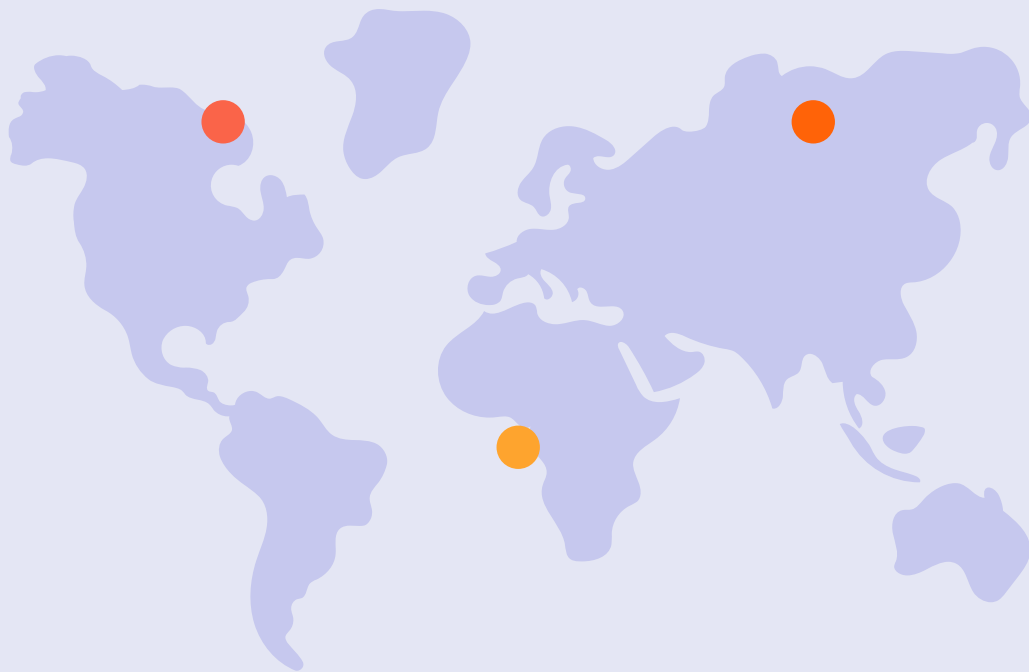
JUPITER

Jupiter is the
biggest planet

Salut!



THIS IS A LANGUAGE MAP



● MARS

Mars is a very cold place

● NEPTUNE

Is the farthest from the Sun

● VENUS

Venus is the second planet

SPEAKERS IN INDIA

● HINDI

Mars is a very cold place

● MARATHI

Mercury is a small planet

● KANNADA

Jupiter is the biggest planet

● TELUGU

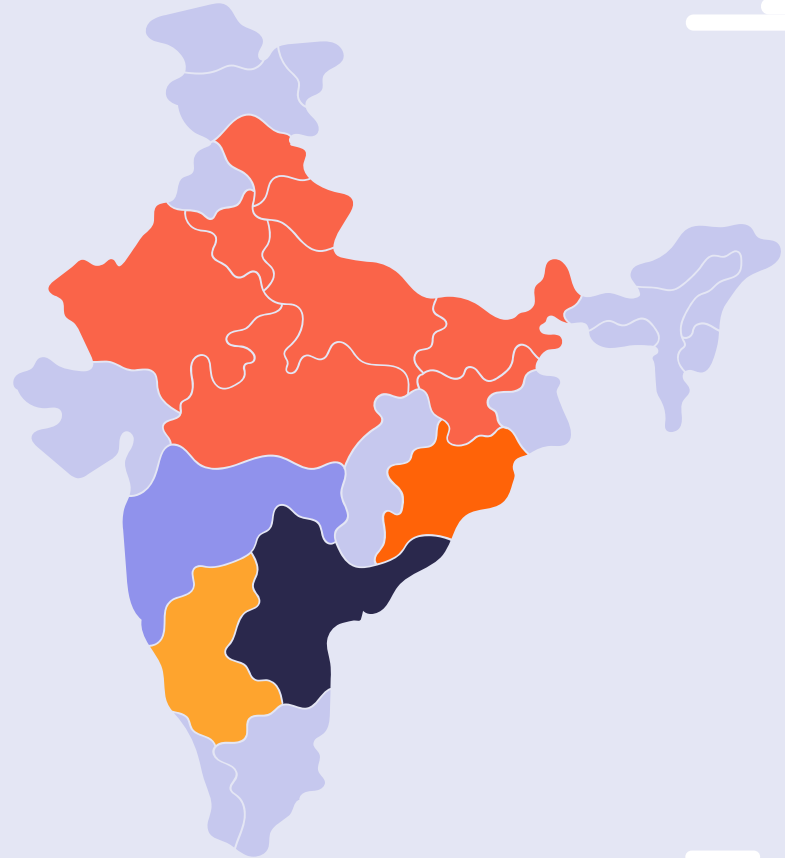
Earth is where we all live on

● ODIA

Neptune is very far away

● OTHERS

Venus has a beautiful name



A TIMELINE ALWAYS WORKS WELL

DAY 1

Mercury is the
smallest planet

DAY 3

Jupiter is the
biggest planet



DAY 2

Mars is a very
cold place

DAY 4

Venus has a
beautiful name



Oi!

6,000

is the approximate number of
languages in the world



2,301

languages
in Asia



2,138

languages
in Africa



1,313

languages
in Oceania

REVIEWING CONCEPTS IS A GOOD IDEA



MERCURY

Mercury is the closest planet to the Sun

VENUS

Venus is the second planet from the Sun

MARS



Despite being red, Mars is a cold place

JUPITER


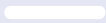
It's the biggest planet in the Solar System

SATURN

Saturn is a gas giant and has rings

NEPTUNE

It's the farthest planet from the Sun



HOW TO SAY “HELLO”...

???



in Korea

???



in Belgium

???



in Angola

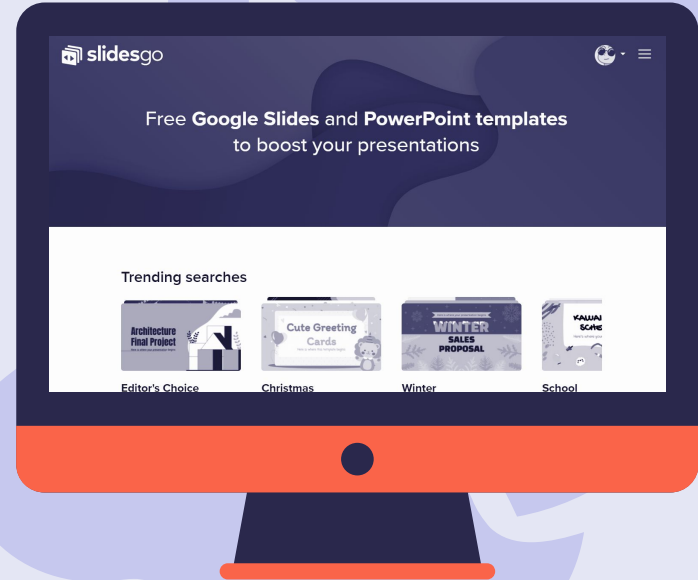
???

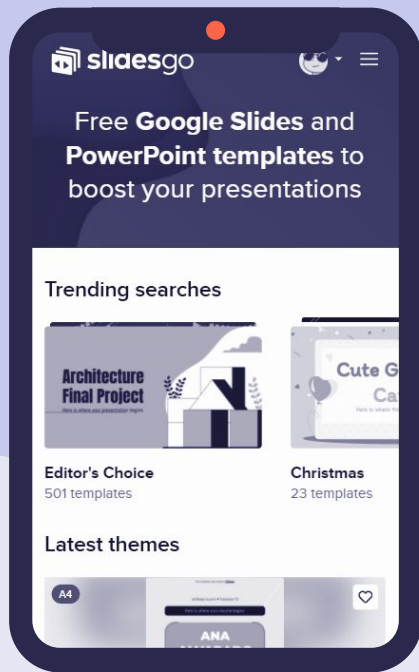


in Russia

DESKTOP SOFTWARE

You can replace the image on the screen with your own work. Just delete this one, add yours and center it properly



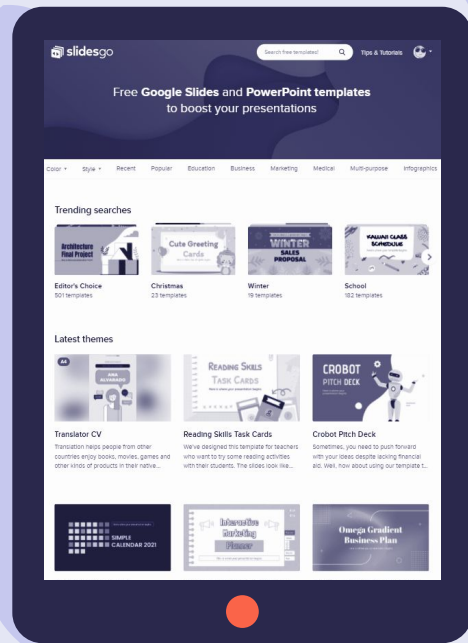


MOBILE SOFTWARE

You can replace the image on the screen with your own work. Just delete this one, add yours and center it properly

TABLET AND MOBILE APP

You can replace the image on
the screen with your own work.
Just delete this one, add yours
and center it properly



OUR TEAM



JOHN DOE

Here you can talk a bit
about this person



HELENA JAMES

Here you can talk a bit
about this person



THANKS

Do you have any questions?

youremail@freepik.com

+91 620 421 838

yourcompany.com

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**

Please keep this slide for attribution



ALTERNATIVE RESOURCES



RESOURCES

ILLUSTRATIONS:

- Language concept background with words
- Different looking people avatar set
- Young old people avatars
- Different people avatars pack
- Collection colorful speech bubbles flat design
- Hand drawn india map infographic

PHOTOS:

- Afroamerican model shouting copy space
- Community young people posing together
- Front view surprised man standing



Instructions for use (free users)

In order to use this template, you must credit [Slidesgo](#) by keeping the Thanks slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>

Instructions for use (premium users)

In order to use this template, you must be a Premium user on [Slidesgo](#).

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.
- Hide or delete the “Thanks” slide and the mention to Slidesgo in the credits.
- Share this template in an editable format with people who are not part of your team.

You are not allowed to:

- Sublicense, sell or rent this Slidesgo Template (or a modified version of this Slidesgo Template).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>

Fonts & colors used

This presentation has been made using the following fonts:

Oswald

(<https://fonts.google.com/specimen/Oswald>)

Roboto

(<https://fonts.google.com/specimen/Roboto>)

#2a284c

#e4e6f4

#9092ec

#fa6449

#fea42e

#ea881c

#ff6308

Storyset

Create your Story with our illustrated concepts. Choose the style you like the most, edit its colors, pick the background and layers you want to show and bring them to life with the animator panel! It will boost your presentation. Check out [How it Works](#).



Pana



Amico



Bro



Rafiki



Cuate

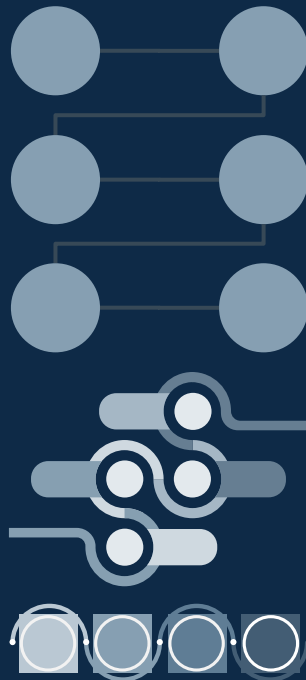
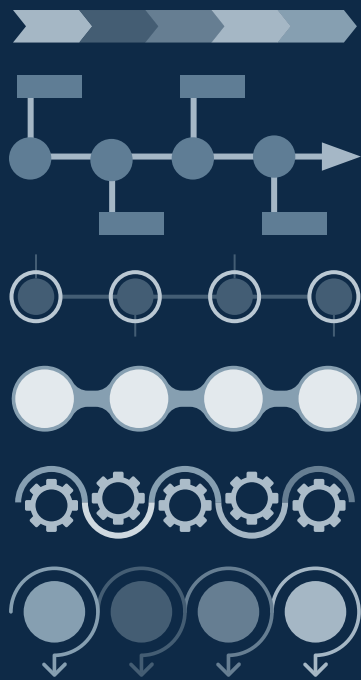
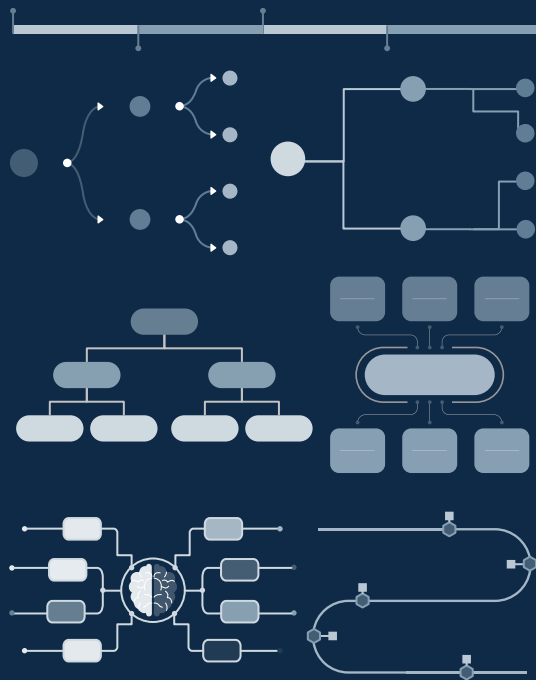
Use our editable graphic resources...

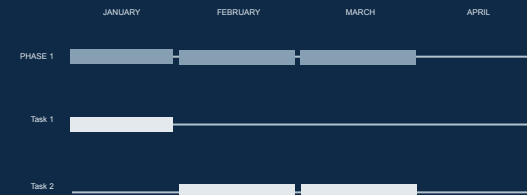
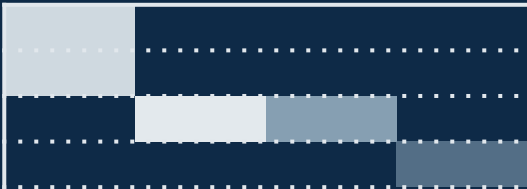
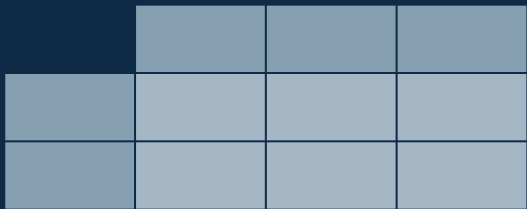
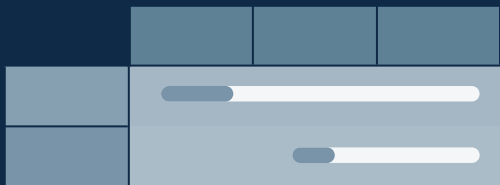
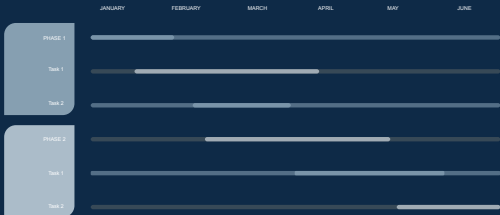
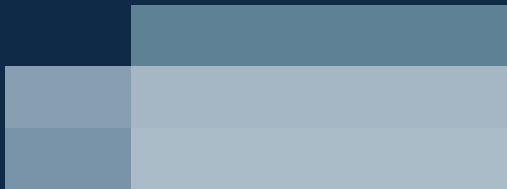
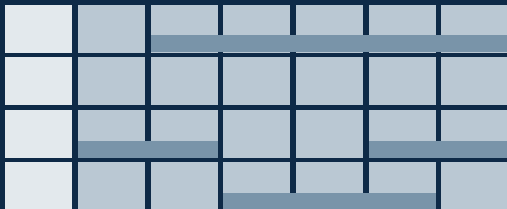
You can easily [resize](#) these resources without losing quality. To [change the color](#), just ungroup the resource and click on the object you want to change. Then, click on the paint bucket and select the color you want.

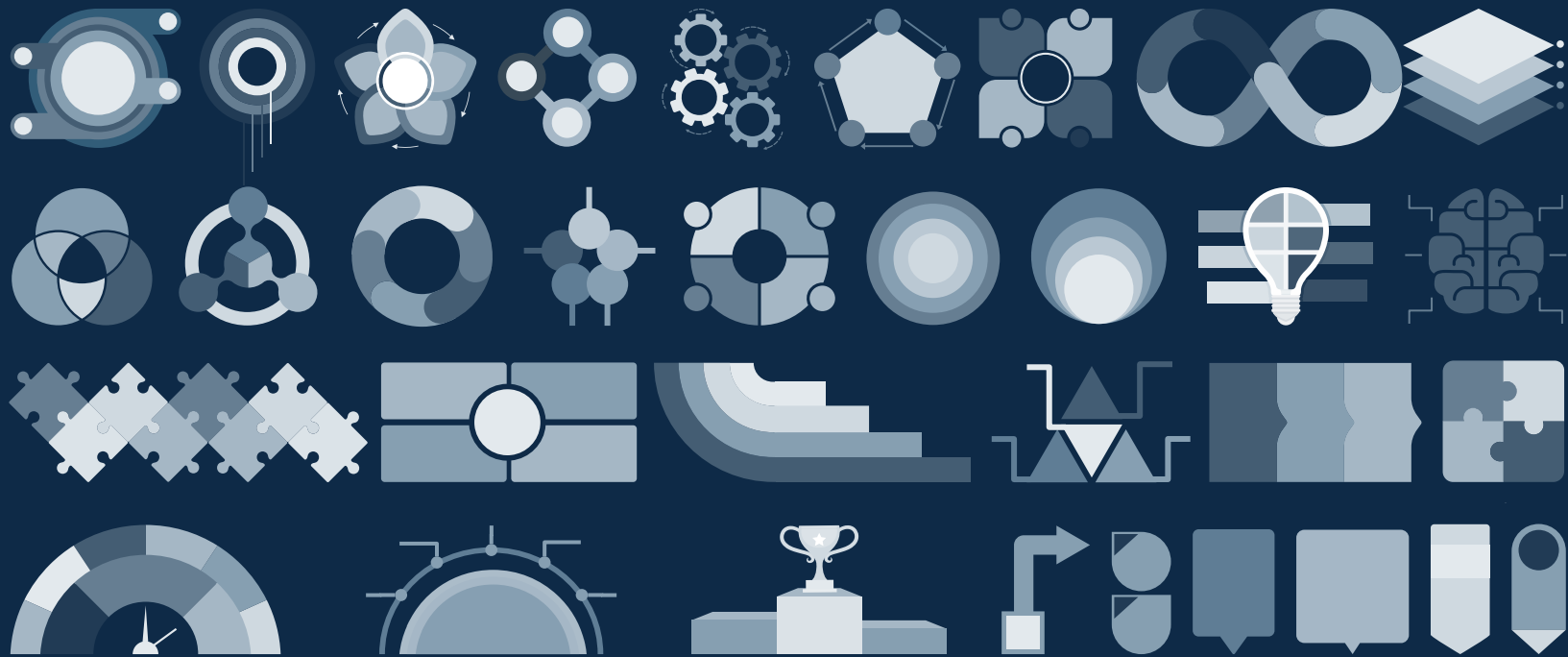
Group the resource again when you're done. You can also look for more [infographics](#) on [Slidesgo](#).

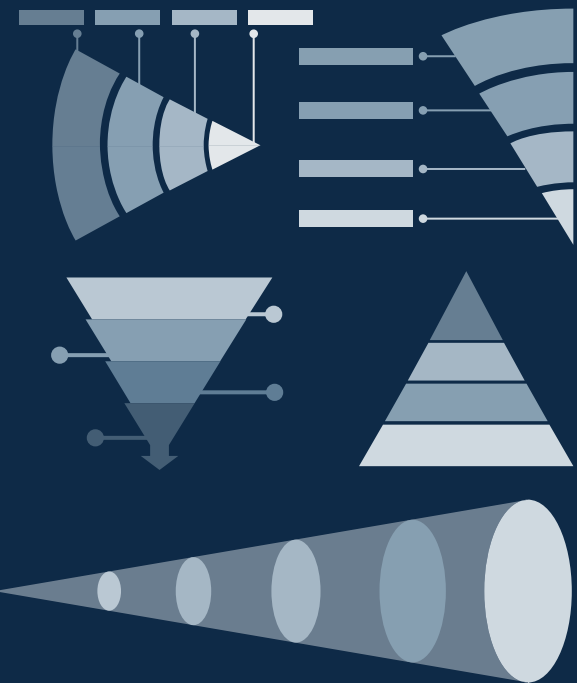
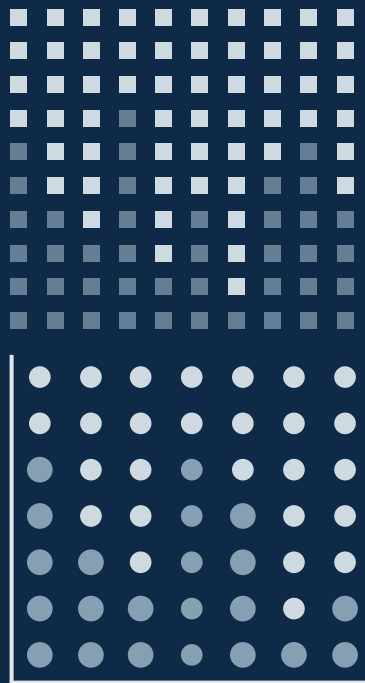












...and our sets of editable icons

You can resize these icons without losing quality.

You can change the stroke and fill color; just select the icon and click on the paint bucket/pen.

In Google Slides, you can also use Flaticon's extension, allowing you to customize and add even more icons.



Educational Icons



Medical Icons



Business Icons



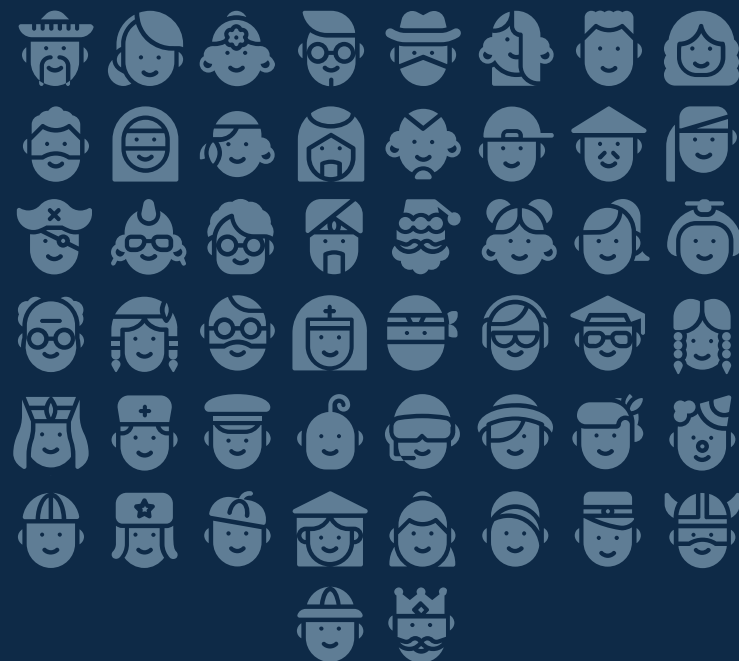
Teamwork Icons



Help & Support Icons



Avatar Icons



[illegible][illegible]

Nature Icons



SEO & Marketing Icons



CONTENTS OF THIS TEMPLATE

Here's what you'll find in this **Slidesgo** template:

- A slide structure based on a multi-purpose presentation, which you can easily adapt to your needs. For more info on how to edit the template, please visit **Slidesgo School** or read our **FAQs**.
- An assortment of illustrations that are suitable for use in the presentation can be found in the **alternative resources** slide.
- A **thanks** slide, which you must keep so that proper credits for our design are given.
- A **resources** slide, where you'll find links to all the elements used in the template.
- **Instructions for use.**
- Final slides with:
 - The **fonts and colors** used in the template.
 - A selection of **illustrations**. You can also customize and animate them as you wish with the online editor. Visit **Storyset** to find more.
 - More **infographic resources**, whose size and color can be edited.
 - Sets of **customizable icons** of the following themes: general, business, avatar, creative process, education, help & support, medical, nature, performing arts, SEO & marketing, and teamwork.

You can delete this slide when you're done editing the presentation.

