A Linguistics-based Approach for Achieving Sentence-level Differential Privacy

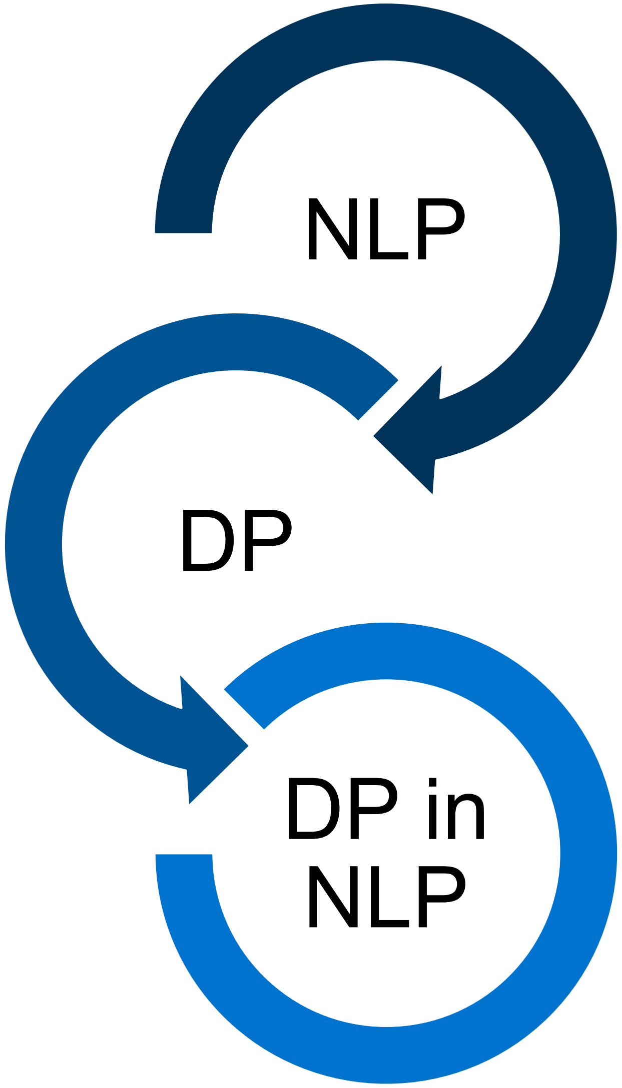
Chaeeun (Joy) Lee

12.02.2024, Bachelor Thesis Kick-Off Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Outline

Motivation



- A critical field within artificial intelligence
- Various applications:

  language translation, sentiment analysis, chatbots, LLM ...

=> Need for effective privacy-preserving techniques in NLP

“Robust **privacy-enhancing technique** offering a mathematically rigorous framework that provides strong privacy guarantees by **introducing controlled noise to individual data points**”

Dwork et al. (2006, "Differential Privacy").

- Challenges to apply DP to unstructured textual data
- Balancing **privacy and preserving meaning & readability**

Motivation

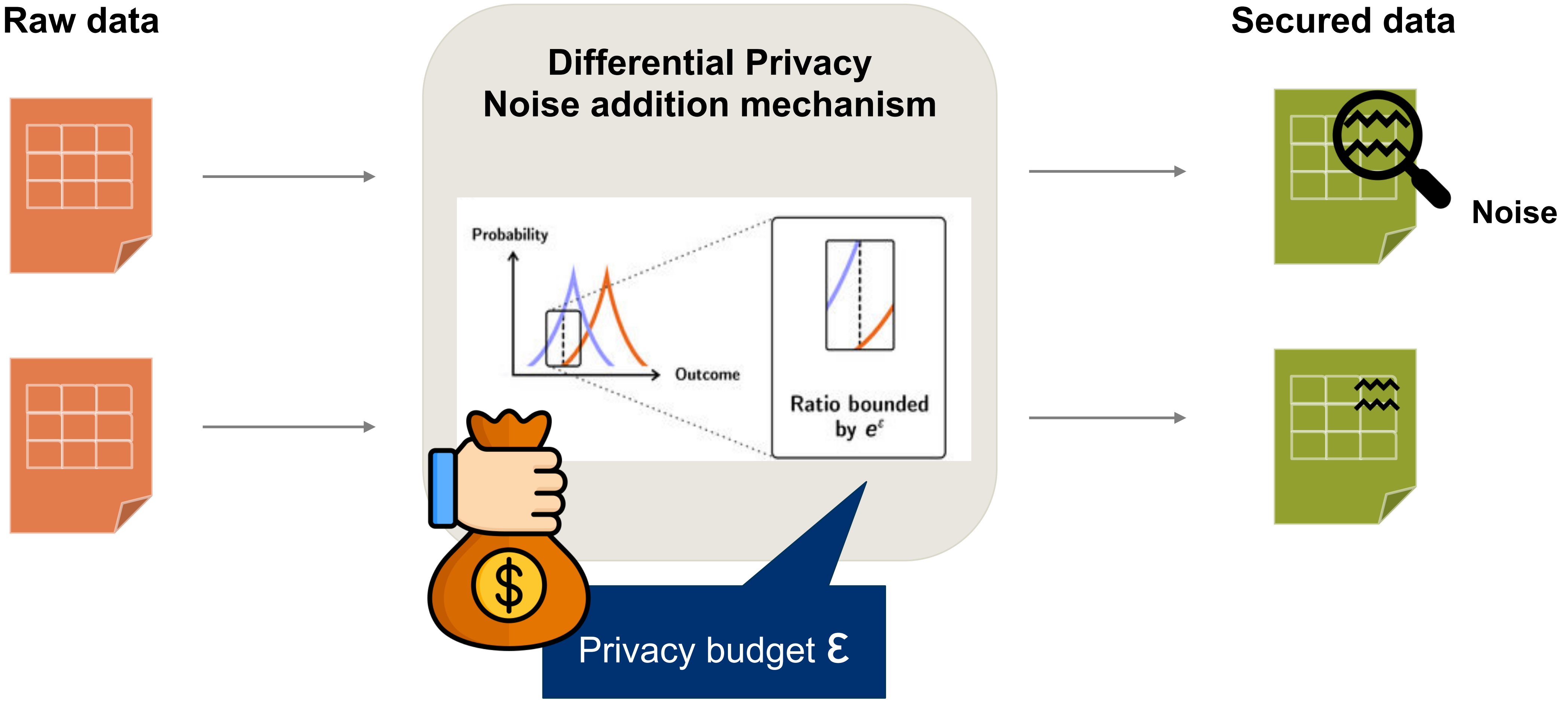


Image : Franzen, Daniel & Nuñez von Voigt, Saskia & Sörries, Peter & Tschorsch, Florian & Müller-Birn, Claudia. (2022). "Am I Private and If So, how Many?" -- Using Risk Communication Formats for Making Differential Privacy Understandable.

Motivation : Conventional approach of applying DP to a Sentence

ε = 1.0

She enjoys reading novels in her cozy, quiet room.

Applied to each individual word in the sentence equally

[She] [enjoys] [reading] [novels] [in] [her] [cozy] [quiet] [room]

↓ 0.1 ↓ 0.1 … ↓ 0.1 ↓ 0.1

[He] [delights] [devouring] [books] [within] [his] [snug] [tranquil] [space]

He delights devouring books within his snug, tranquil space.

Applied to entire embedding of the sentence

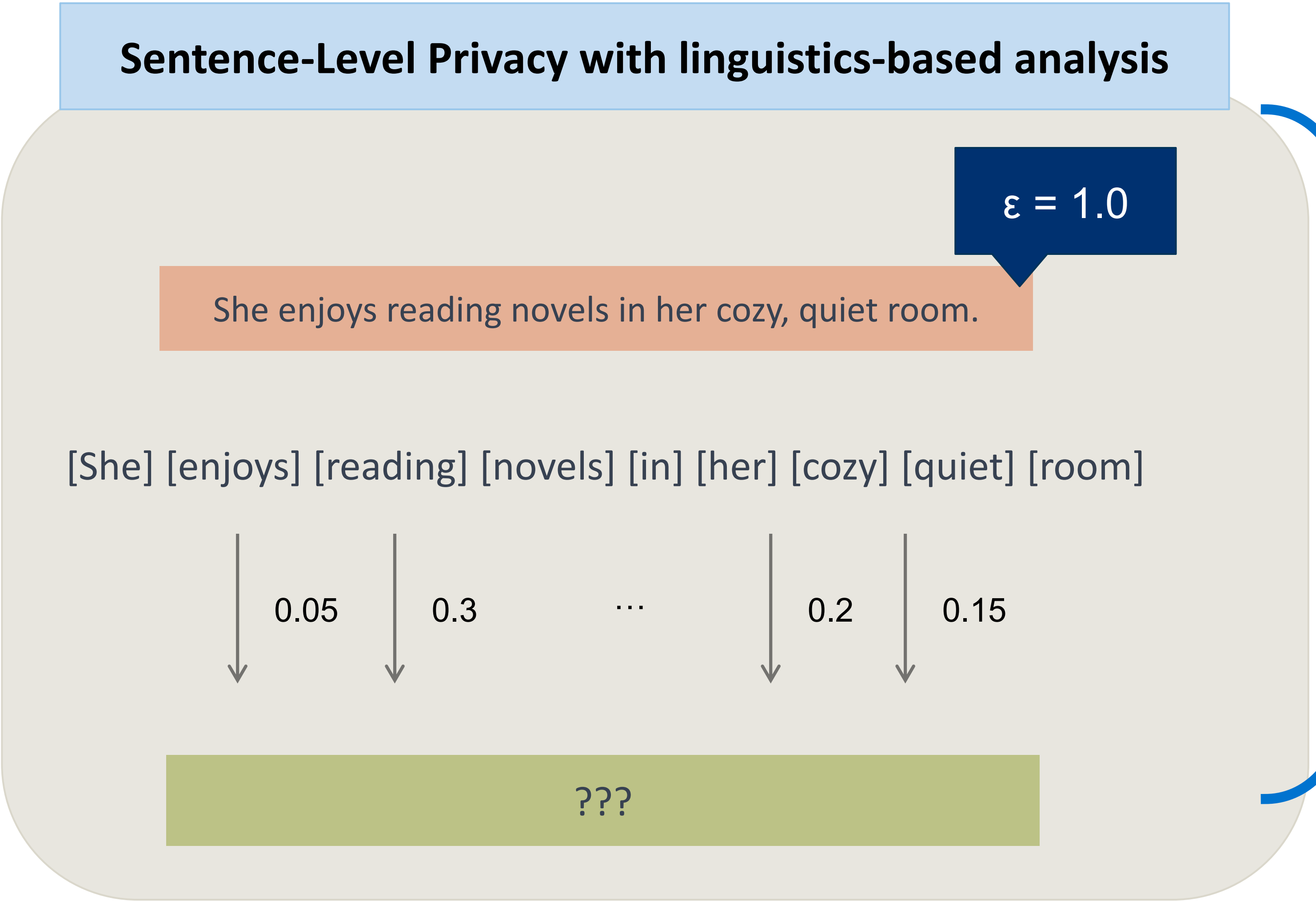
[She enjoys reading novels in her cozy, quiet room.]

↓ 1.0

[People find solace exploring stories in a peaceful environment]

People find solace exploring stories in a peaceful environment.

Motivation : New approach



What is the **intelligent way to distribute the budget** to achieve the sentence-level DP?

Hypothesis :

The higher the informativeness of a word, the greater the likelihood that privacy protection will be necessary.

Research Questions

RQ1

How can DP be effectively applied at the sentence level in NLP, considering the intelligent distribution of privacy budgets for individual words within a sentence?

RQ2

How can the theoretical concepts of sentence-level privacy based on linguistics-based analysis be translated into an implementable framework?

RQ3

How well does the suggested differential privacy approach protect private data while preserving the readability of the text data?

Methodology

Theoretical Research

- Conduct literature review
- Explore linguistics-based methods to calculate informativeness for reasonable distribution of privacy budget across individual words

Implementation

- Design and develop an prototype for the distribution
- Incorporate linguistic models to adjust privacy budget distribution rate to each word

Evaluation

- Analyse utility & privacy evaluation measures comparing to the naive approach
- Evaluation of readability through survey

Expected outcomes

Conceptual Contribution

- Enhanced sentence-level differential privacy, addressing privacy budget challenges at the granularity of individual words within a sentence
- Advancement of theoretical understanding at the intersection of sentence-level in NLP, contributing to broader privacy discourse

Methodological / Practical Contribution

- Establishment of a implementable DP framework for sentence-level privacy, integrating linguistic methods for quantifying word informativeness
- Practical evaluation of the DP framework, offering insights into its effectiveness in protecting sensitive information while maintaining textual coherence in diverse NLP applications
- Suggest useable solution for practical use cases with finite privacy budget

Initial Findings and Progress : Different methods to quantify informativeness

Information content

- Computes the Information Content(IC) value based on synsets (sets of synonyms) in WordNet and the IC of those synsets in various corpora (such as SemCor, Brown, etc.)
- Averages these IC values across different corpora to obtain a single IC score for each word

POS tag

- Assigns score based on part-of-speech (POS) tag
- Used averaged perceptron tagger from NLTK based on the Penn Treebank POS tagset
- Weights is determined by statistic of twitter data but can be customizable by the user

Similarity

- Computes the sentence embedding and modified sentence embedding using a pre-trained Sentence Transformer model (Sentence-BERT)
- Importance measure cosine similarity between embeddings

Name entity recognition

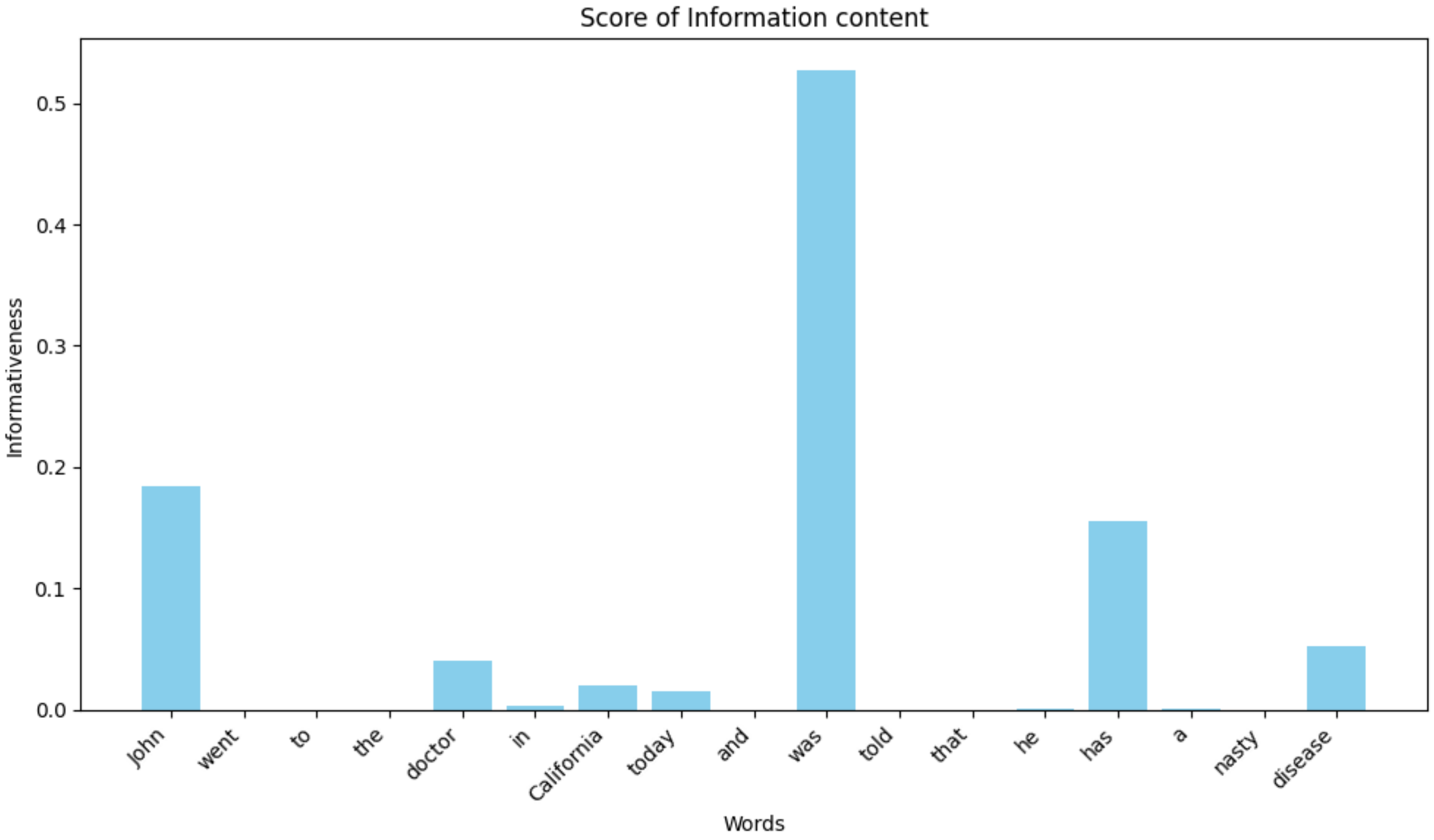
- Uses named entity by a pre-trained spaCy NER model
- If a word is part of a named entity assigns a higher weight

Initial Findings and Progress : Different methods to quantify informativeness

Information content

- Computes the Information Content(IC) value based on synsets (sets of synonyms) in WordNet and the IC of those synsets in various corpora (such as SemCor, Brown, etc.)
- Averages these IC values across different corpora to obtain a single IC score for each word

John went to the doctor in California today and was told that he has a nasty disease

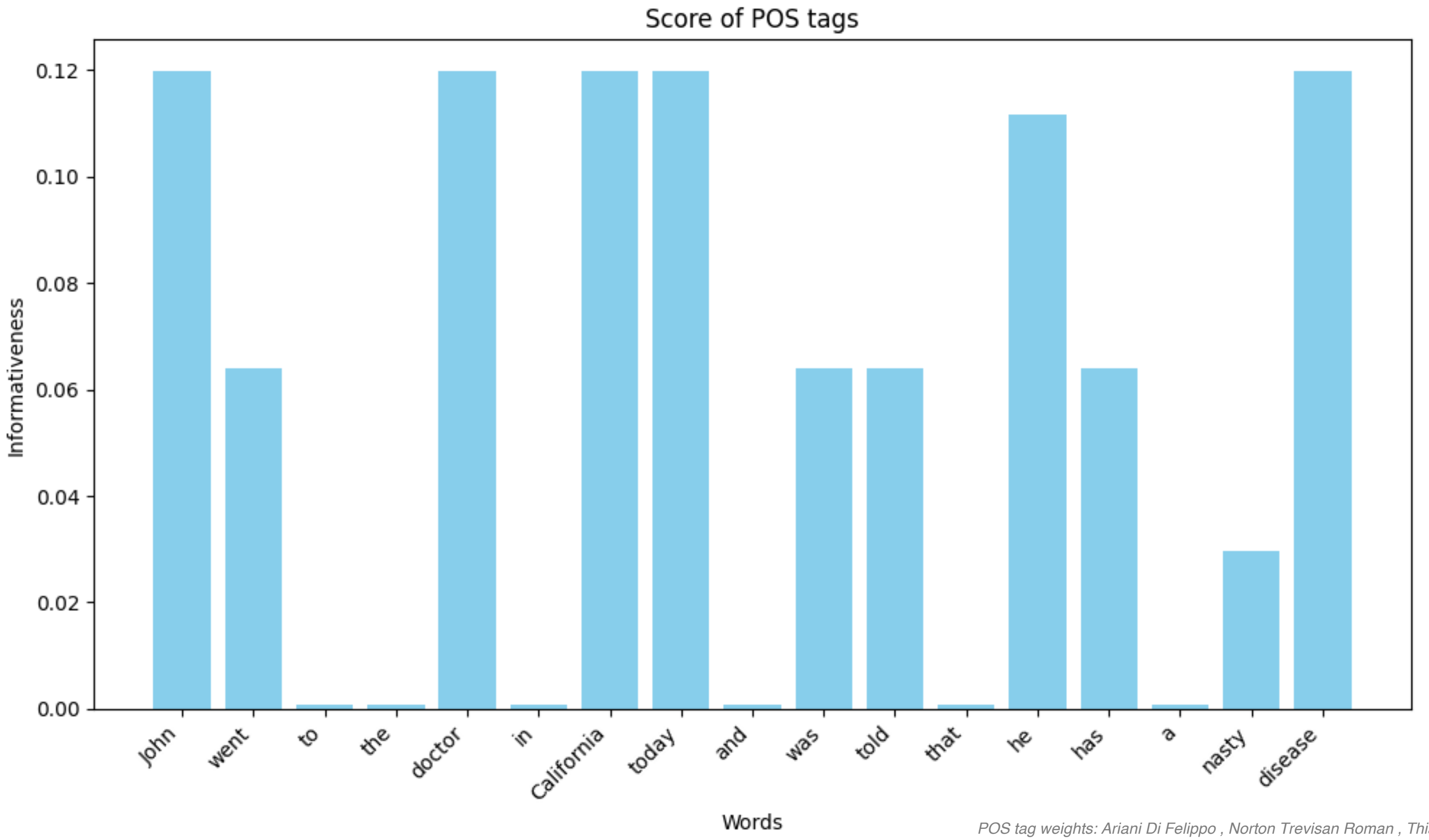


Initial Findings and Progress : Different methods to quantify informativeness

POS tag

- Assigns score based on part-of-speech (POS) tag
- Used averaged perceptron tagger from NLTK based on the Penn Treebank POS tagset
- Weights is determined by statistic of twitter data but can be customizable by the user

John went to the doctor in California today and was told that he has a nasty disease



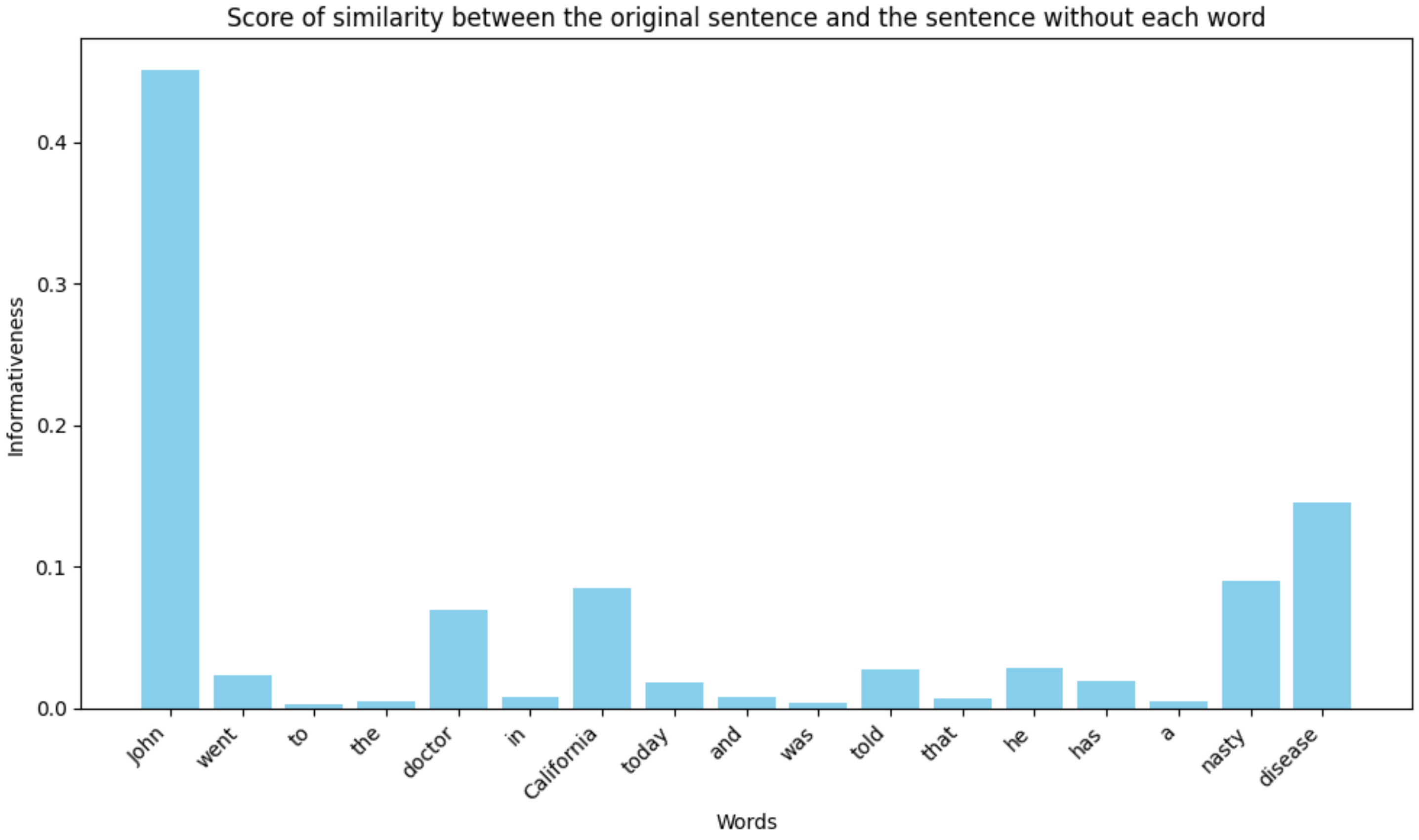
POS tag weights: Ariani Di Felippo , Norton Trevisan Roman , Thiago A. S. Pardo , et al. THE DANTESTOCKS CORPUS: AN ANALYSIS OF THE DISTRIBUTION OF UNIVERSAL DEPENDENCIES-BASED PART OF SPEECH TAGS. TechRxiv. November 28, 2022.

Initial Findings and Progress : Different methods to quantify informativeness

Similarity

- Computes the sentence embedding and modified sentence embedding using a pre-trained Sentence Transformer model (Sentence-BERT)
- Importance measure cosine similarity between embeddings

John went to the doctor in California today and was told that he has a nasty disease

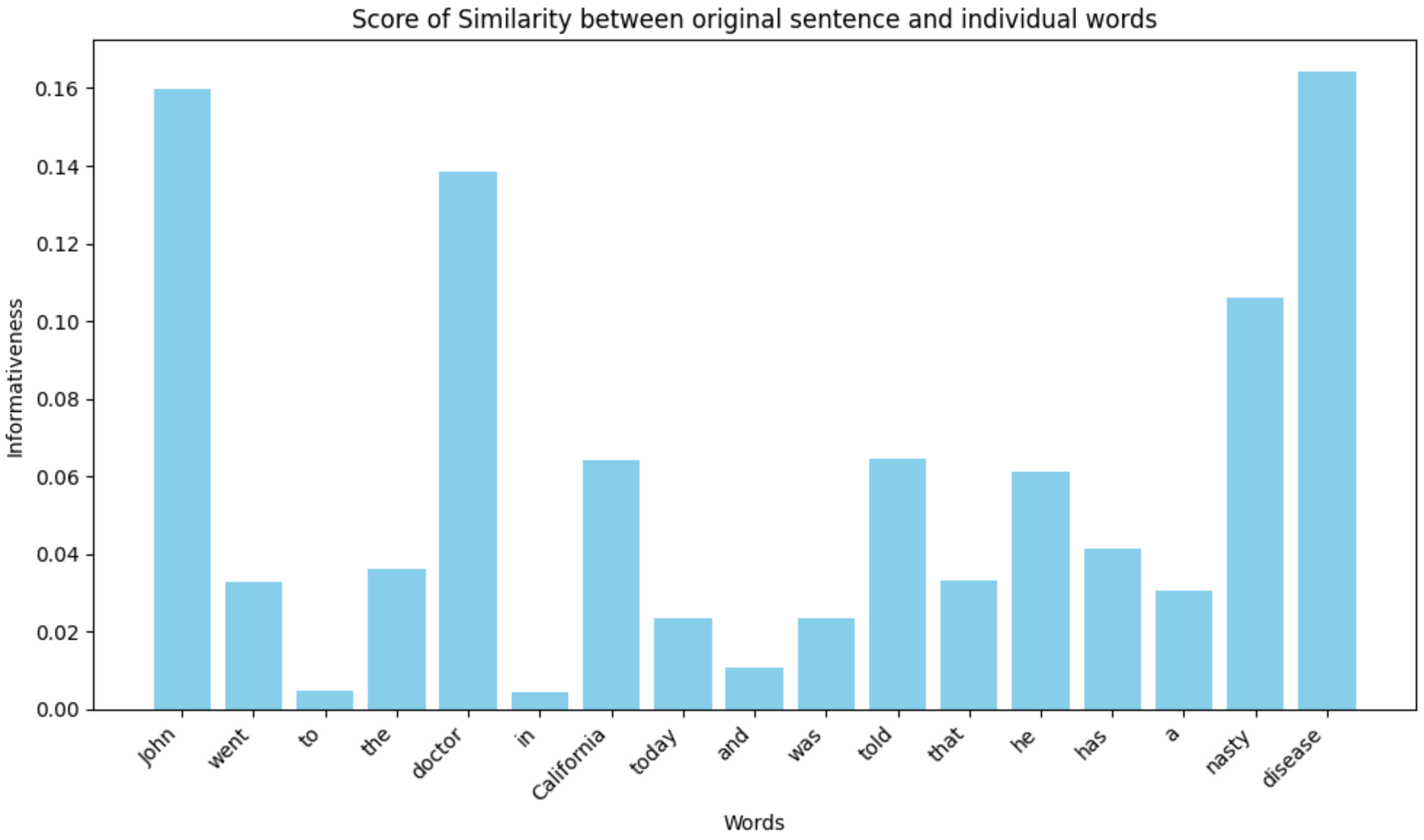


Initial Findings and Progress : Different methods to quantify informativeness

Similarity

- Computes the sentence embedding and modified sentence embedding using a pre-trained Sentence Transformer model (Sentence-BERT)
- Importance measure cosine similarity between embeddings

John went to the doctor in California today and was told that he has a nasty disease

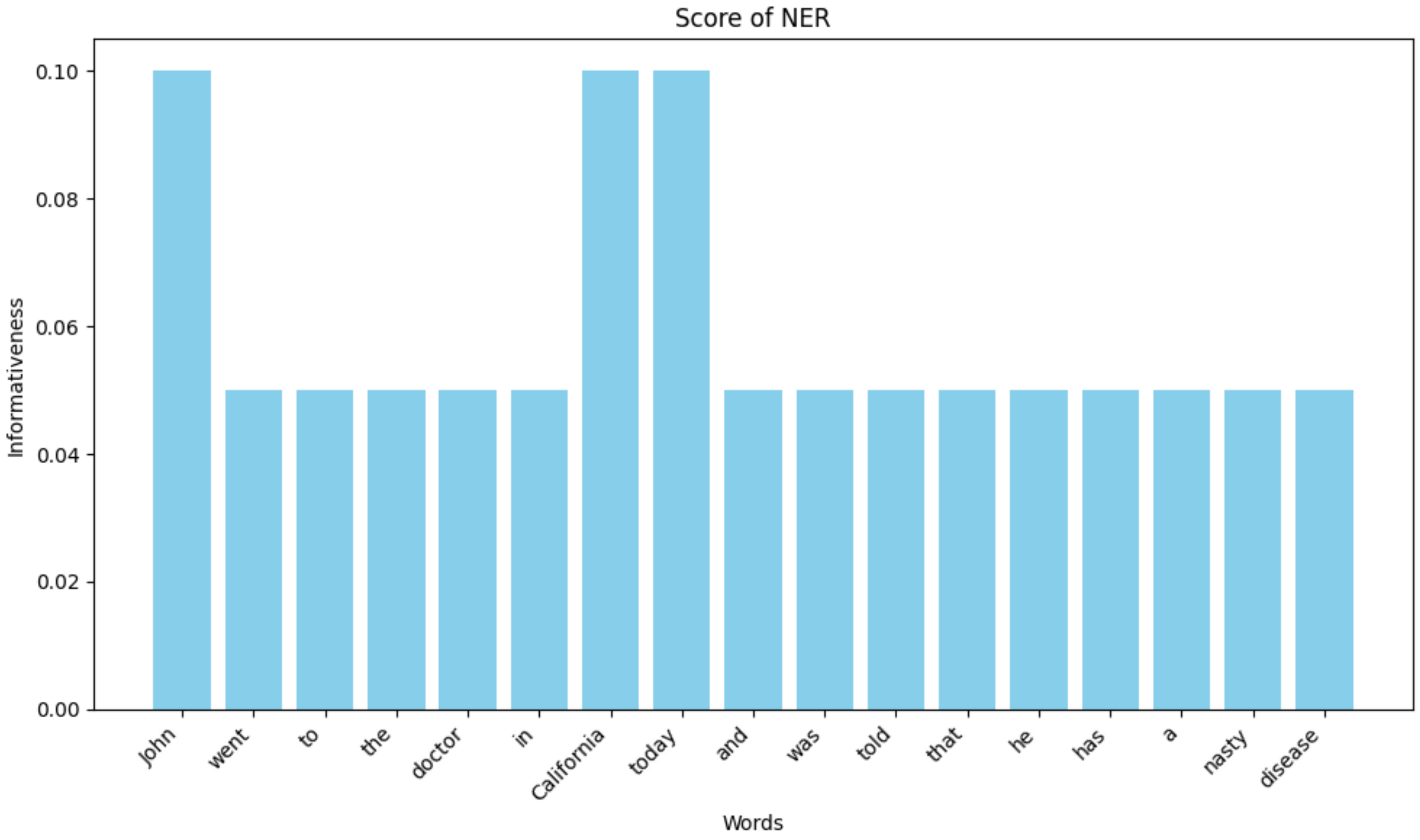


Initial Findings and Progress : Different methods to quantify informativeness

Name entity recognition

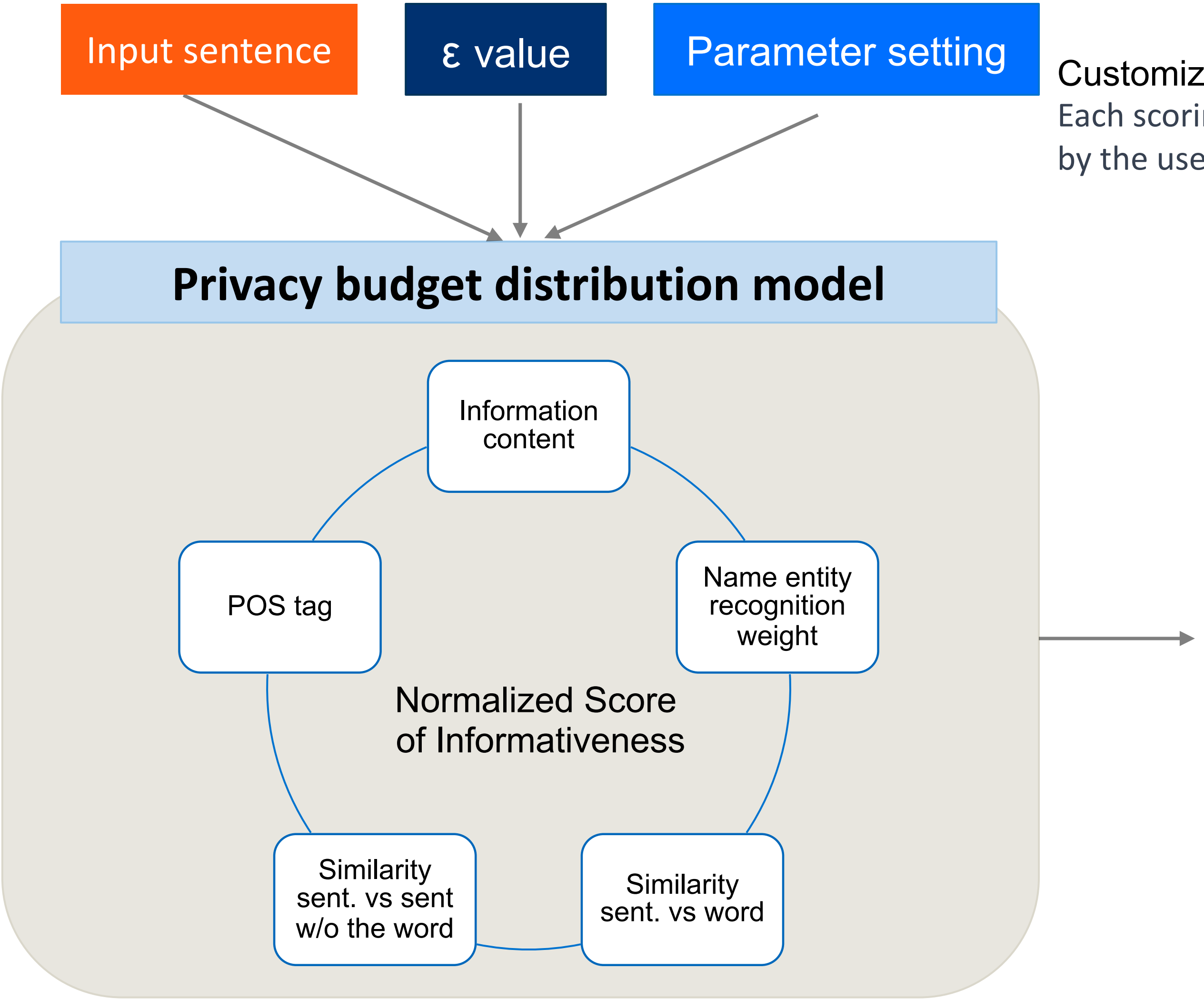
- Uses named entity by a pre-trained spaCy NER model
- If a word is part of a named entity assigns a higher weight

John went to the doctor in California today and was told that he has a nasty disease



Detected Named Entities:
```
John - PERSON
California - GPE
today - DATE
```

Initial Findings and Progress : Model Design



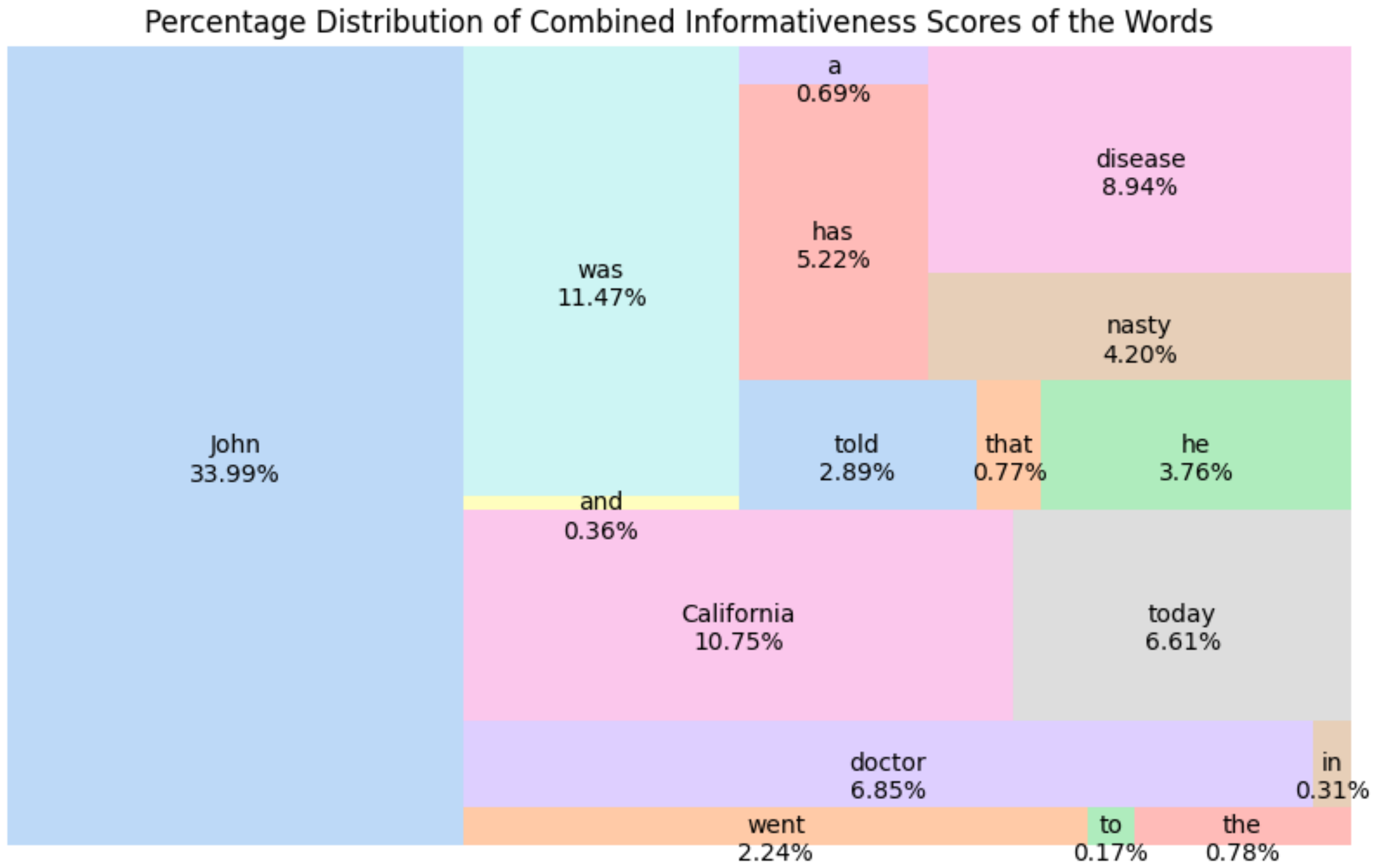
Customizable parameters:
Each scoring method can be selectively enabled or disabled by the user to suit their specific needs and objectives

For each word in a sentence..

- Final combined informativeness scores
- Distributed ε values based on the total privacy budget

Initial Findings and Progress : Initial Result

John went to the doctor in California today and was told that he has a nasty disease



Initial Findings and Progress : Initial Result

John went to the doctor in California today and was told that he has a nasty disease

Naive approach

Kevin came at the friend in CA Today
& WAS contacted THAT
he got really lovely disease

VS

**Suggested approach
with budget distribution**

Ed went see my dentist in Maryland,
but was diagnosed that
he got this thyroid migraine

Next Steps

Utility & Privacy Evaluation

Compare results of proposed methods against the naive approach to assess effectiveness.

Evaluate utility and privacy using relevant measures to quantify improvements.

Readability Analysis

Design survey questions to gather data on readability.

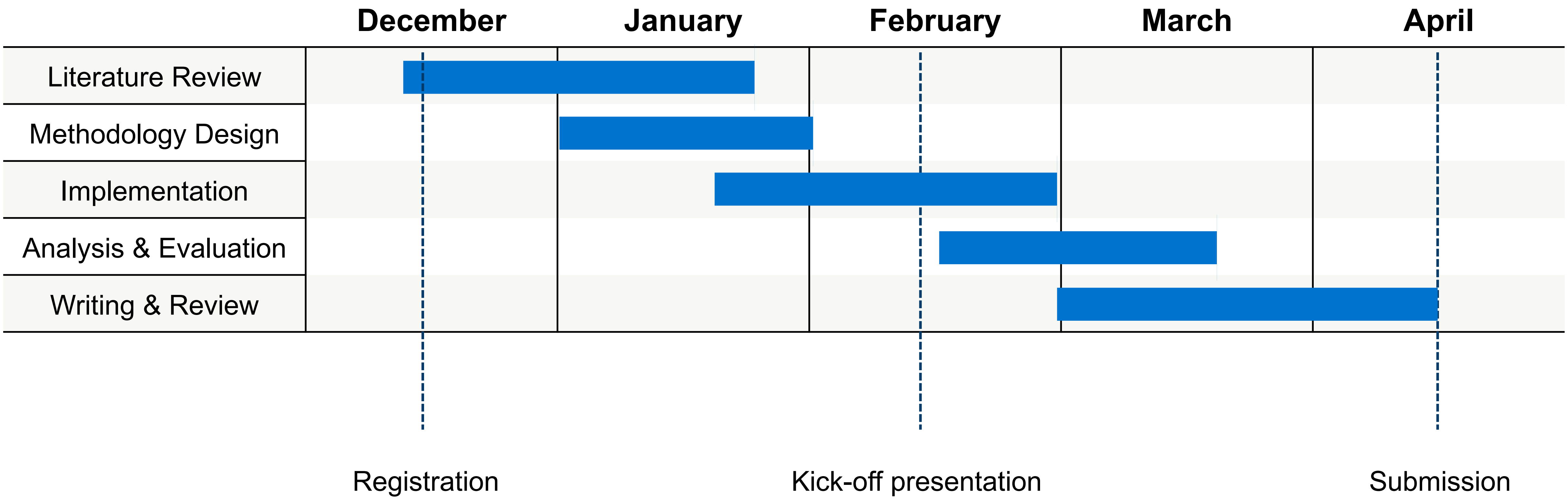
Analyse survey responses to assess the readability of the proposed methods.

Concluding into Paper

Summarize the literature review, methodology, and planned evaluation.

Highlight how addressing identified gaps and challenges contributes to the research field.

Timeline


December
January
February
March
April
Literature Review
Methodology Design
Implementation
Analysis & Evaluation
Writing & Review
Registration
Kick-off presentation
Submission

Chaeeun (Joy) Lee

Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.0000
chaeeun.joy.lee@tum.de
wwwmatthes.in.tum.de