Empirical study of Terms of Service

Maciej Kasztelanic
Sergio Carcamo

Data and Data processing

- Over 80 Terms of Service from SaaS companies
- Removing common but different parts like

  company name, emails, urls, phones, address
- Removing law specific parts like roman letters, annotations

Embeddings

Word embeddings - pretrained BERT model

vector of length 512 for every company

Sentence embeddings - SentenceTransformer

vector with no max length for every company

Testing statistical differences in similarities

BERT: little differentiation across categories

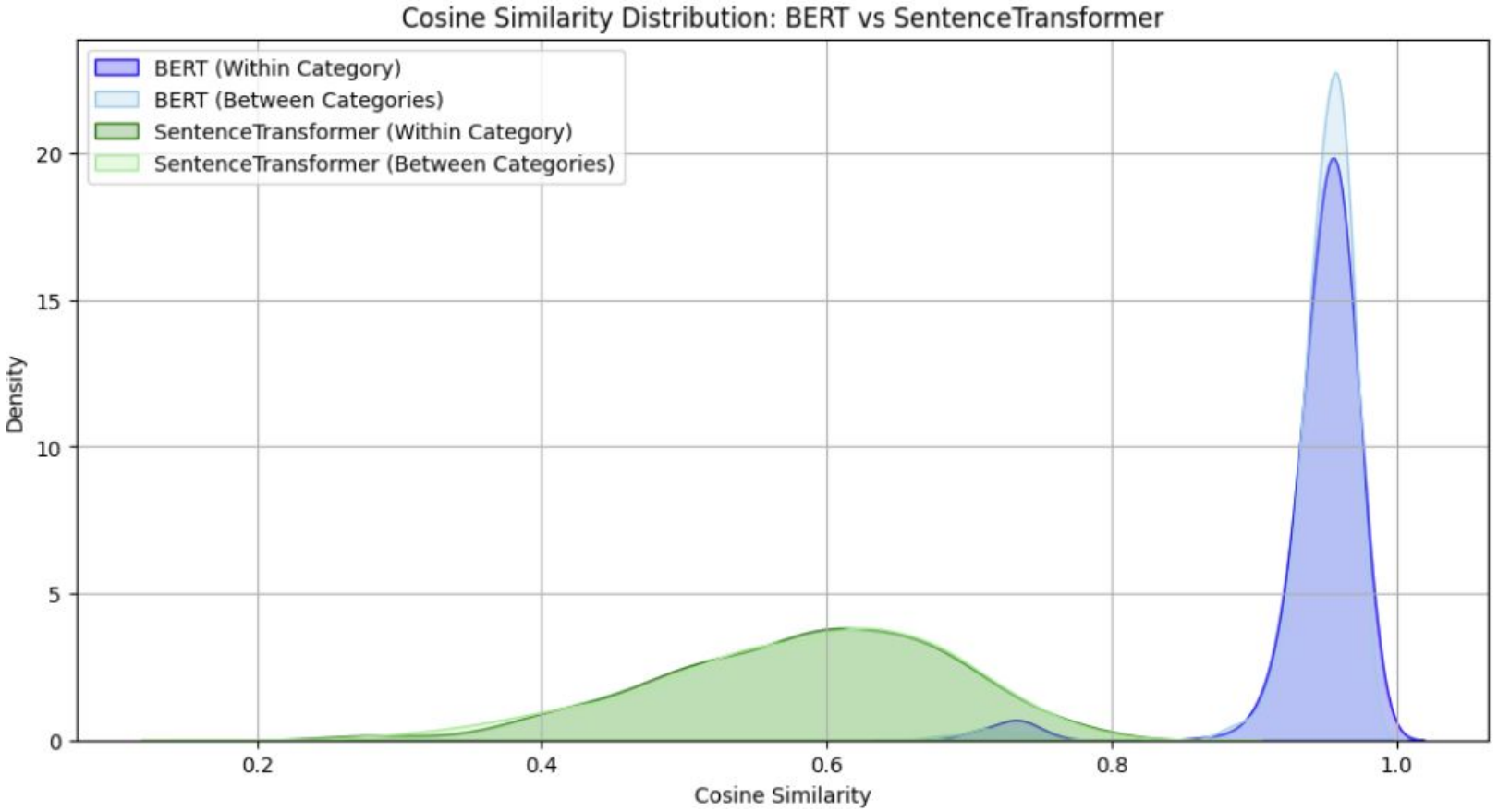
within-group (0.944) | between-group (0.946)

SentenceTransformer: minimal distinction between

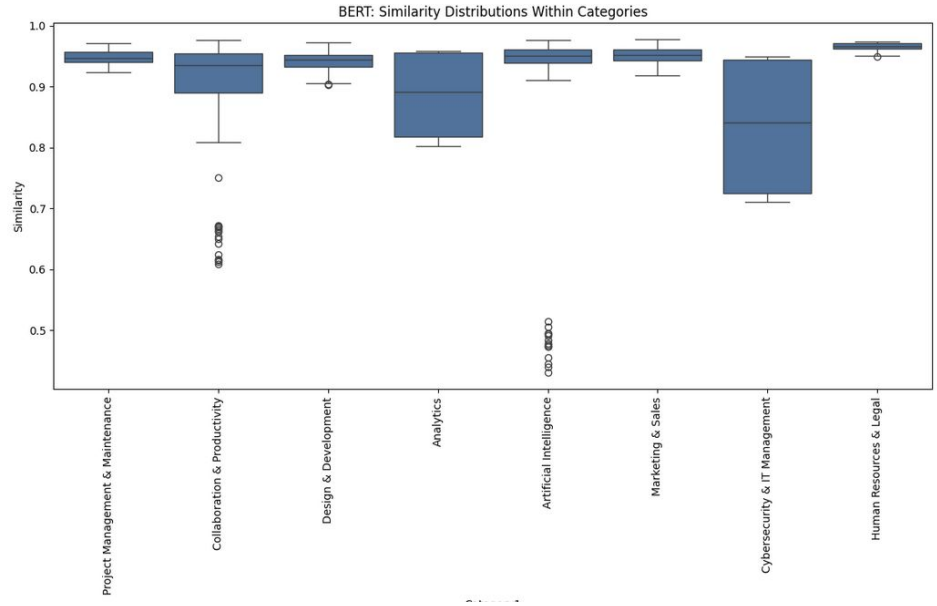
within-group (0.583) | between-group (0.582)

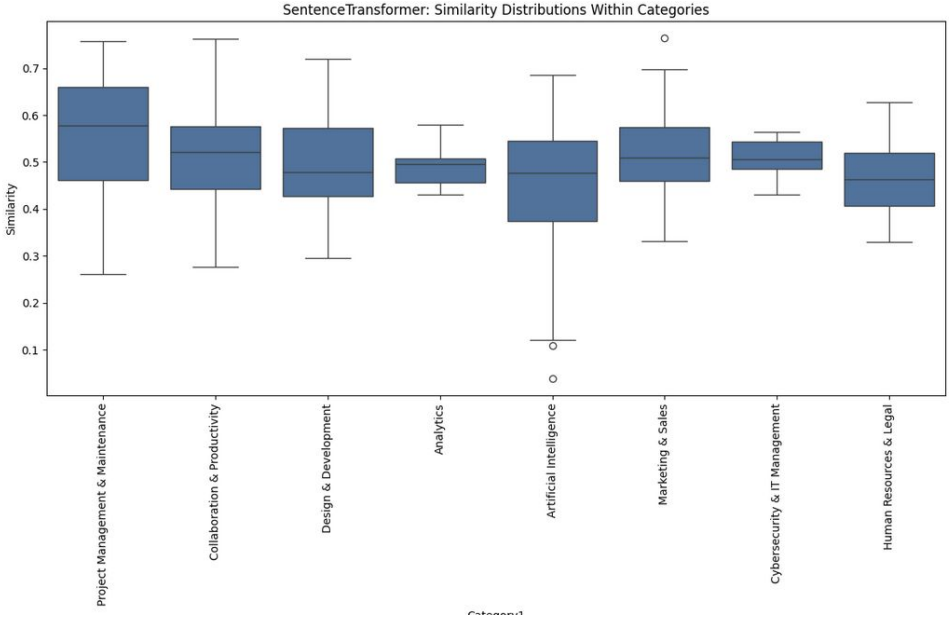
ANOVA	Kruskal - Wallis
BERT: p = 0.37,	BERT: p = 0.88
SentenceTransformer: p = 0.76	SentenceTransformer: p = 0.95

Similarity within and between categories

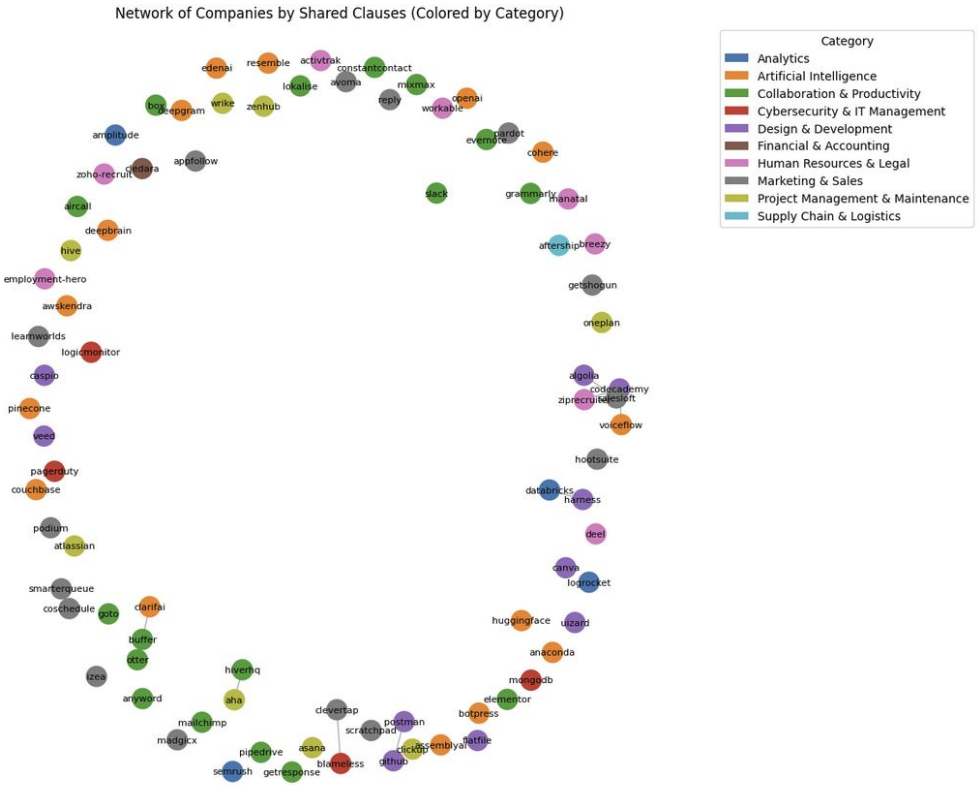


Within category similarity





Network analysis of similarities



connection is when at least 10% or more of the n-grams in the ToS of one company overlap with those in another.

Topic modelling

For topic modelling of the terms of service we followed this process:

Split the document into sections, using section headings

Remove urls, company names, and other stop words

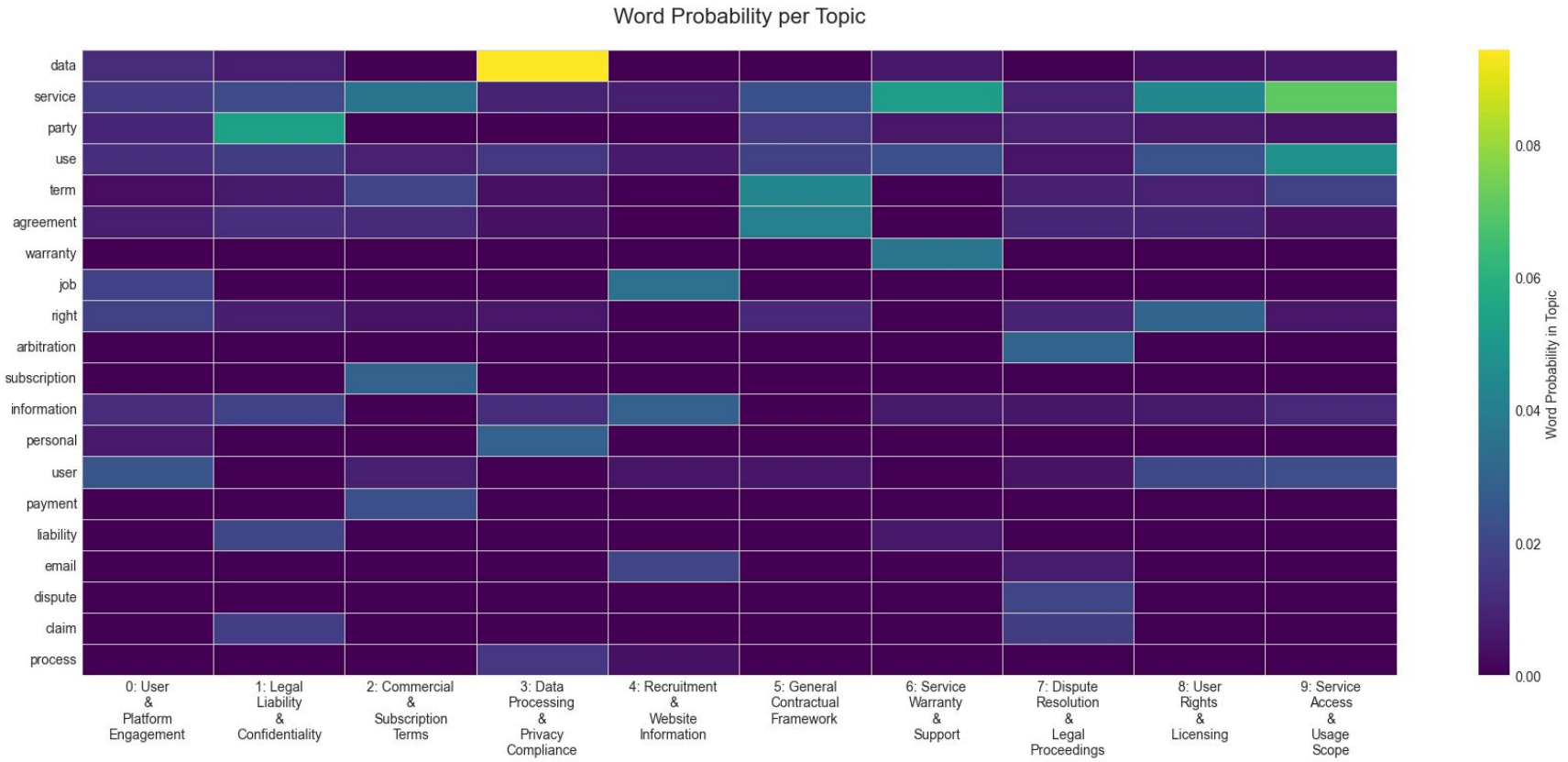
Apply lemmatization

Perform Topic modeling using LDA

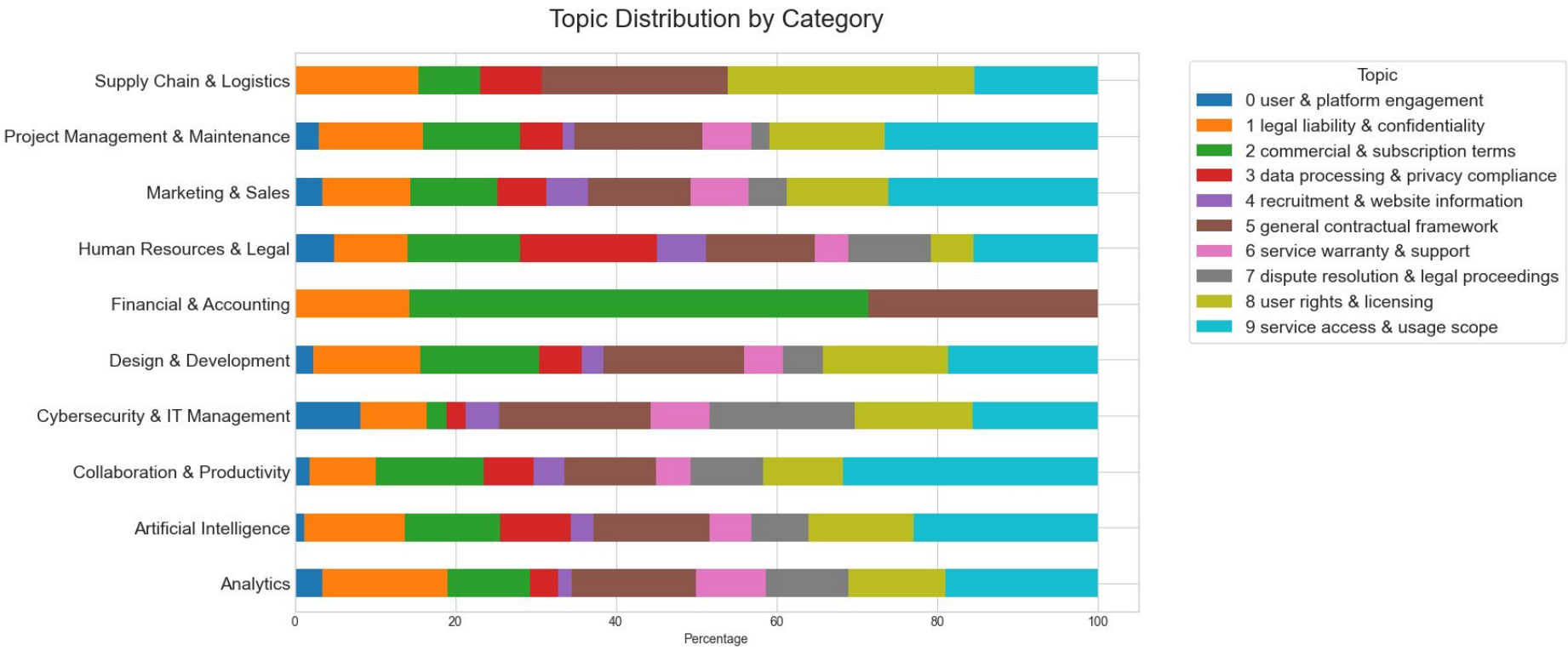
Topic modelling

- **Topic 0:** User and Platform Engagement
- **Topic 1:** Legal Liability and Confidentiality
- **Topic 2:** Commercial and Subscription Terms
- **Topic 3:** Data Processing and Privacy Compliance
- **Topic 4:** Recruitment and Website Information
- **Topic 5:** General Contractual Framework
- **Topic 6:** Service Warranty and Support
- **Topic 7:** Dispute Resolution and Legal Proceedings
- **Topic 8:** User Rights and Licensing
- **Topic 9:** Service Access and Usage Scope

Topic modelling



Topic modelling



Distribution by category