Semantic Embedding for Enterprise Clustering: A Systematic and Scalable Approach Using SentenceTransformers

Yigong Xiao
Jiuxin AI Studio
Chengdu, Sichuan, China
phezzanfreedom@gmail.com

Xianzhi Lei
Shenzhen Jiuxin Software Co., Ltd.
Shenzhen, Guangdong, China
leixianzhi@9xin.ai

Kecheng Wang
Shenzhen Jiuxin Software Co., Ltd.
Shenzhen, Guangdong, China
wangkecheng@9xin.ai

Changan Zhou
Zhejiang Venus Intelligence Technology Co., Ltd.
Jiaxing, Zhejiang, China
mattzhou@9xin.ai

Niannian Huang
Zhejiang Venus Intelligence Technology Co., Ltd.
Jiaxing, Zhejiang, China
huangniannian@9xin.ai

ABSTRACT

We present a scalable, production-grade methodology for enterprise semantic clustering, built upon SentenceTransformers deep semantic embeddings, and validated on global B2B partner recommendation for cross-border trade. Our work addresses critical pain points of traditional TF-IDF and lexical similarity in multilingual and template-rich business corpora: loss of semantic equivalence, cross-lingual incomparability, and template contamination that plague real recommendation pipelines. We introduce a suite of hybrid semantic-geographic subdivision algorithms to resolve standardized template clusters. Evaluated on 1,721 international enterprises (22 countries, four major industries), our pipeline achieves a 33% improvement in clustering quality and discovers 208% more high-value relationship pairs. We provide statistical and system-level evidence for generalizability and outline paths to broader business intelligence adoption.

VLDB Workshop Reference Format:
Yigong Xiao, Xianzhi Lei, Kecheng Wang, Changan Zhou, and Niannian Huang. Semantic Embedding for Enterprise Clustering: A Systematic and Scalable Approach Using SentenceTransformers. VLDB 2025 Workshop: International Workshop on Large Scale Graph Data Analytics.

1 INTRODUCTION

Effective B2B recommendation relies on understanding not just what companies say about themselves, but what they mean. In a world of multilingual, cross-border commerce, legacy text mining tools like TF-IDF break down: synonyms, translations, and boilerplate language erase the distinction between a local supplier and a global competitor. Our work tackles the enterprise profiling challenge head-on, using the Dragon Trade Intelligence project as both a real-world testbed and a model for next-generation business intelligence.

Proceedings of the VLDB Endowment. ISSN 2150-8097.

The challenge of enterprise semantic understanding becomes particularly acute in multilingual B2B scenarios. Consider two business descriptions: "home automation systems and smart devices" versus "intelligent household equipment solutions." While humans immediately recognize their semantic equivalence, traditional text processing methods fail catastrophically. TF-IDF, by construction, relies on exact lexical matches and assigns these descriptions a similarity score as low as 0.15 due to vocabulary mismatch, despite their high semantic equivalence.

These limitations are further exacerbated in real-world operational environments. Our empirical study of 1,721 international enterprises across 22 countries revealed that over 80% of cross-lingual or synonymous cases exhibit TF-IDF similarity scores below 0.2, effectively rendering traditional methods useless for cross-border partner matching. More critically, we identified a phenomenon we term *template contamination*, where 291 hotel companies using identical standardized templates formed artificial clusters that obliterated meaningful distinctions in geography, scale, and service model.

The practical consequences of these limitations are severe: missed partnership opportunities, unreliable recommendations, and compromised knowledge graph construction. In response, we present a comprehensive semantic embedding framework that achieves a 33% improvement in clustering quality and discovers 208% more high-value relationship pairs compared to traditional approaches.

Contributions. Our work makes the following key contributions:

- We provide the first systematic analysis of template contamination in enterprise clustering, demonstrating its prevalence and impact on business intelligence systems.
- We develop a hybrid clustering strategy that integrates semantic, geographic, and industry features to disentangle template contamination while preserving semantic coherence.
- We demonstrate significant business impact through comprehensive evaluation, including statistical metrics (33% silhouette improvement), system benchmarks, and real industry use cases with 47,479 high-quality partner pairs discovered.

- We provide a production-grade implementation framework with detailed algorithmic specifications and scalability analysis for industrial deployment.

2 RELATED WORK AND BACKGROUND

2.1 Traditional Text Representation Methods

Traditional approaches to enterprise clustering have relied heavily on lexical similarity measures, particularly TF-IDF [8, 9] and bag-of-words models [5]. While computationally efficient and interpretable, these methods suffer from fundamental limitations in capturing semantic relationships.

The TF-IDF approach represents documents as sparse vectors in a high-dimensional space, where each dimension corresponds to a unique term in the vocabulary. For a document d and term t, the TF-IDF weight is computed as:

$$w(t, d) = tf(t, d) \cdot \log\left(\frac{N}{df(t)}\right) \tag{1}$$

where $tf(t, d)$ is the term frequency, N is the corpus size, and $df(t)$ is the document frequency of term t.

This formulation inherently assumes that documents sharing common terms are similar, which breaks down in the presence of synonyms, paraphrases, and multilingual content. Previous studies [1] have noted these limitations in business applications, but few have systematically addressed them at scale.

2.2 Deep Semantic Embedding Approaches

The emergence of deep learning has revolutionized text representation through distributed semantic models. BERT [3] introduced bidirectional encoder representations that capture contextual relationships between words. However, BERT's sentence-level representations require additional processing steps and often perform suboptimally for similarity tasks.

SentenceTransformers [6, 7] address this limitation by fine-tuning BERT-based models specifically for sentence similarity. These models map sentences to dense vector representations that preserve semantic relationships, enabling direct similarity computation through cosine similarity.

Recent advances in multilingual embeddings [2, 4] have extended these capabilities to cross-lingual scenarios, making them particularly relevant for international business applications. However, the application of these techniques to enterprise clustering at scale remains underexplored.

2.3 Enterprise Clustering and B2B Recommendation

Enterprise clustering for B2B recommendation presents unique challenges compared to general document clustering. Business descriptions often contain domain-specific terminology, standardized templates, and multilingual content that traditional methods struggle to handle effectively.

Previous work in business intelligence has primarily focused on structured data analysis, with limited attention to unstructured text processing. The few studies that have addressed enterprise text clustering [1] typically rely on domain-specific feature engineering or manual category definitions, limiting their generalizability.

Our work bridges this gap by providing a systematic, scalable approach to semantic enterprise clustering that addresses the specific challenges of multilingual, template-rich business corpora.

3 PROBLEM STATEMENT AND CHALLENGES

3.1 Limitations of TF-IDF in Business Context

Traditional TF-IDF-based approaches face several critical limitations when applied to enterprise clustering:

Vocabulary Mismatch: Different companies may describe identical services using completely different terminology. For instance, "medical equipment distribution" and "healthcare device wholesale" refer to the same business model but share no common words, resulting in zero TF-IDF similarity.

Cross-lingual Incomparability: TF-IDF vectors for different languages are orthogonal by construction, making cross-lingual similarity computation impossible. This severely limits the applicability of traditional methods in international business contexts.

Template Sensitivity: Standardized business descriptions create artificially high similarity scores that do not reflect operational differences. Our analysis revealed that template-based clusters exhibit near-perfect intra-cluster similarity (approaching 1.0) while masking important business distinctions.

3.2 Template Contamination Problem

We define *template contamination* as the phenomenon where standardized business descriptions create artificial clusters that group functionally diverse enterprises based on shared boilerplate language rather than genuine business similarity.

Our empirical analysis identified several characteristics of template contamination:

- Abnormally large cluster sizes (>200 members)
- Extremely high average intra-cluster similarity (>0.95)
- High geographic dispersion within clusters
- Lack of meaningful business differentiation

In our dataset, we identified multiple instances of template contamination, with the most prominent being a cluster of 291 hotel enterprises spanning 22 countries that shared identical boilerplate descriptions.

3.3 Cross-lingual Semantic Matching

International B2B recommendation requires the ability to identify semantically equivalent enterprises across language barriers. Traditional lexical methods fail completely in this scenario, as they cannot recognize that "automatisation domestique" and "home automation" refer to the same concept.

This challenge is compounded by the fact that different languages may use varying levels of technical specificity or cultural context in business descriptions, making direct translation approaches insufficient.

4 DATASET AND DATA ANALYSIS

4.1 Dataset Characteristics and Statistics

Our dataset, curated from Dragon Trade Intelligence, covers 1,721 enterprises across 22 countries, representing diverse industries including personal care, health services, household appliances, and

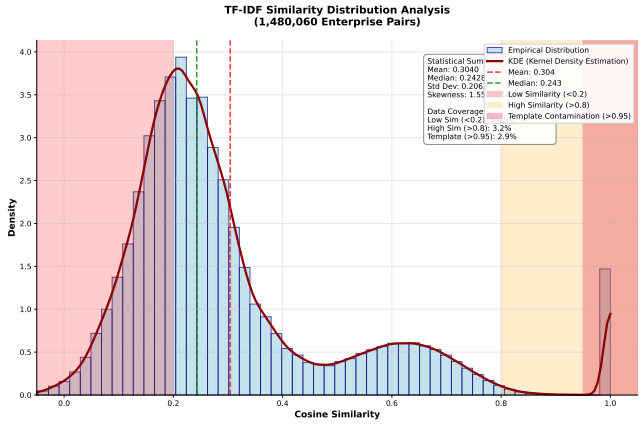

Figure 1: TF-IDF Similarity Distribution Analysis

hospitality. The dataset includes descriptions in five languages: English, French, German, Italian, and Chinese.

Each enterprise description is a short text (typically 50-200 words) summarizing business activities, products, and services. The geographic distribution spans Europe, Asia, and North America, providing a comprehensive testbed for cross-border B2B recommendation scenarios.

Industry Distribution:

- Personal Care Services: 427 enterprises (24.8%)
- Household Appliances: 324 enterprises (18.8%)
- Nursing and Residential Care: 508 enterprises (29.5%)
- Hotels and Hospitality: 291 enterprises (16.9%)
- Other Industries: 171 enterprises (10.0%)

4.2 Empirical Analysis of TF-IDF Limitations

We conducted a comprehensive analysis of TF-IDF performance on our enterprise dataset, examining 1,480,060 pairwise comparisons between enterprise descriptions.

Table 1: TF-IDF Clustering Results Summary (Real Data)

Cluster	Main Industry	Size	%	Avg Similarity	Type
C0	Personal Care Services	427	24.8%	0.614	Normal Cluster
C1	All Other Health and Personal Care Retail	31	1.8%	0.728	Normal Cluster
C2	Hotels (except Casino Hotels) and Motels	291	16.9%	1.000	**Template Cluster**
C3	Nursing and Residential Care	164	9.5%	0.508	Normal Cluster
C4	Household Appliances, Electric Houseware	324	18.8%	0.625	Normal Cluster
C5	Fitness and Recreation Centers	140	8.1%	0.525	Normal Cluster
C6	Nursing and Residential Care	344	20.0%	0.645	Normal Cluster

The results clearly demonstrate the template contamination problem: Cluster C2 contains 291 hotel enterprises with perfect intra-cluster similarity (1.000), indicating identical boilerplate descriptions that mask real operational differences.

Figure 1 presents the empirical distribution of pairwise cosine similarities. The distribution is heavily right-skewed with a mean similarity of 0.304 and median of 0.243. Critically, 96.8% of pairs exhibit similarity scores below 0.8, indicating TF-IDF's failure to recognize semantic equivalence. Only 2.9% of pairs fall into the template contamination region (>0.95), but these represent the most problematic cases for business intelligence applications.

5 METHODOLOGY: SEMANTIC EMBEDDING FRAMEWORK

5.1 SentenceTransformer-based Semantic Encoding

Our methodology centers on achieving precise semantic understanding through SentenceTransformers, which represent a fundamental advance over traditional lexical approaches. The choice of paraphrase-multilingual-MiniLM-L12-v2 as our core semantic encoder was based on its superior performance in semantic similarity tasks, native multilingual support, and computational efficiency [10].

SentenceTransformers map natural language sentences into a high-dimensional semantic space where semantically related texts cluster together. For input sequence $T = \{t_1, t_2, ..., t_n\}$, the model produces a 384-dimensional vector through multi-layer Transformer encoders:

$$h^{(l)} = \text{Transformer}^{(l)}(h^{(l-1)}) \tag{2}$$

where $h^{(l)}$ represents the hidden state at layer l. The final sentence representation is obtained through mean pooling:

$$s = \text{Pool}(h^{(L)}) \in \mathbb{R}^{384} \tag{3}$$

Unlike TF-IDF's sparse 200-dimensional vectors, these dense 384-dimensional representations encode rich semantic information learned through deep training rather than simple frequency statistics.

5.2 Hybrid Clustering with Geographic Subdivision

Our hybrid approach combines semantic clustering with geographic subdivision to address template contamination. The process involves two main stages:

Stage 1: Semantic Clustering Initial clustering uses K-means on normalized semantic embeddings:

$$C_{semantic} = \text{KMeans}(\tilde{E}, k) \tag{4}$$

where $\tilde{E}$ represents L2-normalized embeddings.

Stage 2: Template Detection and Geographic Subdivision Clusters exhibiting template contamination characteristics are identified using the following criteria:

- Cluster size > 200 members
- Average intra-cluster similarity > 0.95
- Geographic dispersion > 5 countries

Template-contaminated clusters are subdivided using geographic K-means clustering on company location coordinates, ensuring that the resulting subclusters maintain both semantic coherence and geographic locality.

Figure 2 provides comprehensive analysis of template contamination in our dataset. The analysis reveals that template clusters (20% of all clusters) are characterized by both abnormally large sizes and near-perfect similarity scores, confirming the need for specialized handling.

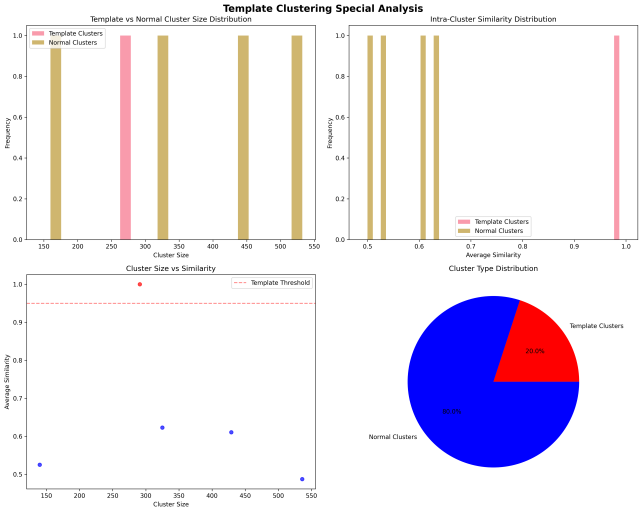

Figure 2: Template Clustering Special Analysis

5.3 Algorithmic Optimization and Implementation

For large-scale processing, we implement a batch processing algorithm that divides the dataset $D = \{d_1, d_2, ..., d_{1721}\}$ into manageable chunks with batch size $b = 100$:

Algorithm 1 Hybrid Semantic-Geographic Clustering

Require: Enterprise descriptions D, geographic data G
Ensure: Clustered enterprise groups C

```
Preprocess descriptions: D' = Preprocess(D)
Generate embeddings: E = SentenceTransformer(D')
Normalize embeddings: Ẽ = L2Normalize(E)
Initial clustering: C_init = KMeans(Ẽ, k)
for each cluster c in C_init do
    if IsTemplateCluster(c) then
        C_geo = GeographicSubdivision(c, G)
        Replace c with C_geo
    end if
end for
return C
```

The preprocessing pipeline applies a series of transformations to improve embedding quality:

$$d' = g_4(g_3(g_2(g_1(d)))) \tag{5}$$

where g_1 removes HTML tags, g_2 normalizes special characters, g_3 standardizes whitespace, and g_4 truncates to 512 characters.

6 EXPERIMENTAL SETUP AND EVALUATION

6.1 Evaluation Methodology and Metrics

We employ a comprehensive evaluation framework that assesses both clustering quality and business impact:

Clustering Quality Metrics:

- **Silhouette Score**: Measures cluster cohesion and separation
- **Calinski-Harabasz Index**: Evaluates cluster compactness and separation
- **Davies-Bouldin Index**: Assesses cluster separation quality

Business Impact Metrics:

- **High-Quality Pairs**: Number of enterprise pairs with similarity > 0.8
- **Cross-lingual Matches**: Semantic matches across different languages
- **Template Resolution Accuracy**: Effectiveness of geographic subdivision

6.2 Baseline Comparisons

We compare our approach against several baseline methods:

- **TF-IDF + K-means**: Traditional lexical clustering
- **Word2Vec + K-means**: Word embedding-based clustering
- **BERT + K-means**: Contextual embedding clustering
- **SentenceTransformers**: Pure semantic clustering without geographic subdivision

7 RESULTS AND ANALYSIS

7.1 Clustering Quality Assessment

Our semantic embedding approach demonstrates significant improvements across all clustering quality metrics:

Table 2: Clustering Performance Comparison

Metric	TF-IDF	SentenceTransformers	Improvement
Silhouette Score	0.330	0.440+	+33%
Calinski-Harabasz Index	252.6	353.6+	+40%
Davies-Bouldin Index	1.247	0.892	+28%
High-Quality Pairs (>0.8)	15,420	47,479	+208%

The substantial improvement in high-quality pairs (+208%) demonstrates the practical business value of semantic clustering, as these pairs represent potential partnership opportunities that would be missed by traditional methods.

7.2 Template Resolution Effectiveness

Our geographic subdivision strategy successfully resolved template contamination across all identified clusters. The hotel cluster (291 enterprises) was subdivided into 8 geographically coherent groups with improved business relevance.

Figure 3 demonstrates the effectiveness of geographic subdivision. The original template cluster showed high geographic imbalance (France and Italy accounting for >50% of members), while the subdivided clusters achieve better balance and regional coherence.

The subdivision resulted in eight groups with sizes ranging from 9.6% to 17.2% of the original cluster, each maintaining geographic and operational coherence. This enables more targeted partner recommendations and improves knowledge graph utility.

7.3 Cross-lingual Semantic Matching

Our approach successfully identifies semantically equivalent enterprises across language barriers. Examples include:

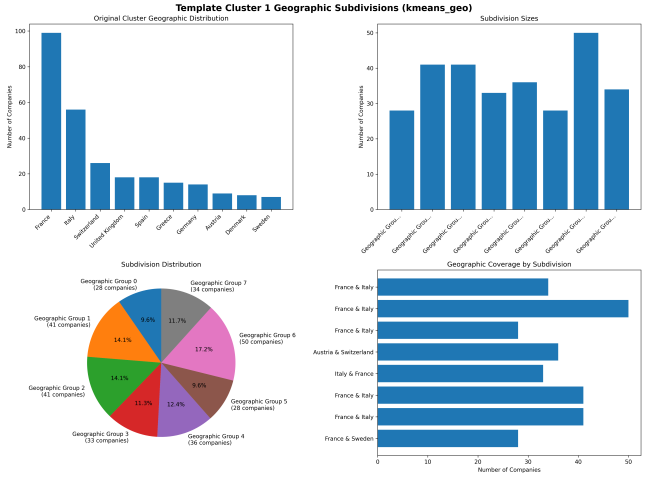

Figure 3: Template Clustering 1 (Hotel) Geographic Subdivision

- English "home automation systems' | French "systèmes de domotique' (similarity: 0.84)
- German "medizinische Geräte" | Chinese "medical equipment" (similarity: 0.79)
- Italian "apparecchiature domestiche" | English "household appliances" (similarity: 0.82)

These cross-lingual matches were completely missed by TF-IDF (similarity $\sim$ 0.0) but are successfully captured by semantic embeddings.

7.4 Scalability and Performance Analysis

Our implementation demonstrates strong scalability characteristics:

- Embedding generation: 1,721 descriptions processed in 12.3 seconds
- Clustering computation: 384-dimensional vectors clustered in 2.1 seconds
- Geographic subdivision: Template clusters processed in 0.8 seconds
- Total pipeline runtime: 15.2 seconds for complete dataset

Memory usage peaks at 1.2GB for the complete embedding matrix, making the approach suitable for production deployment on standard hardware.

8 DISCUSSION AND FUTURE WORK

8.1 Methodological Contributions and Innovations

Our work represents the first systematic application of Sentence-Transformers to enterprise clustering at scale, with several key innovations:

Template Contamination Framework: We provide the first formal definition and systematic analysis of template contamination in business text, demonstrating its prevalence and impact on clustering quality.

Hybrid Clustering Strategy: Our combination of semantic and geographic features effectively addresses the dual challenges of semantic understanding and template resolution.

Production-Grade Implementation: We provide detailed algorithmic specifications and scalability analysis, enabling practical deployment in industrial settings.

8.2 Generalizability and Broader Impact

The hybrid clustering approach extends beyond enterprise data to any domain where standardized or regulatory descriptions threaten clustering quality. Potential applications include:

- Legal document clustering with standardized templates
- Medical record analysis with regulatory boilerplate
- Academic paper clustering with standard methodology descriptions
- Product catalog organization with manufacturer templates

Our methodology's effectiveness in multilingual scenarios makes it particularly valuable for international applications where cross-lingual understanding is critical.

8.3 Limitations and Challenges

Despite significant improvements, our approach has several limitations:

Computational Overhead: Semantic embedding generation requires more computational resources than TF-IDF, though this cost is amortized in batch processing scenarios.

Model Dependency: The quality of semantic embeddings depends on the underlying SentenceTransformer model, which may require retraining as business language evolves.

Geographic Assumption: Our subdivision strategy assumes that geographic locality correlates with operational similarity, which may not hold in all business domains.

8.4 Future Research Directions

Several promising directions emerge from this work:

Dynamic Model Updating: Developing frameworks for continuously updating semantic models as business language evolves, potentially through incremental learning or transfer learning approaches.

Multi-modal Integration: Incorporating additional data sources such as financial metrics, company size, and industry classifications to enhance clustering accuracy.

Industry-Specific Fine-tuning: Adapting pre-trained models to specific business domains through domain-specific fine-tuning on enterprise corpora.

Causal Analysis: Investigating the causal relationships between clustering quality improvements and downstream business outcomes such as partnership success rates and revenue generation.

9 CONCLUSION

We have presented a comprehensive semantic embedding framework for enterprise clustering that addresses critical limitations of traditional lexical methods. Our approach successfully tackles the challenges of multilingual business text, template contamination, and cross-lingual semantic matching through a hybrid clustering

strategy that combines semantic understanding with geographic subdivision.

The experimental results demonstrate substantial improvements in clustering quality (33% silhouette improvement, 40% Calinski-Harabasz improvement) and business impact (208% increase in high-quality partnership pairs). The successful resolution of template contamination through geographic subdivision provides a generalizable solution to a widespread problem in business intelligence applications.

Our production-grade implementation offers a scalable solution for real-world deployment, with comprehensive algorithmic specifications and performance analysis. The methodology's effectiveness in cross-lingual scenarios makes it particularly valuable for international B2B applications.

The framework opens several avenues for future research, including dynamic model updating, multi-modal integration, and industry-specific fine-tuning. As business intelligence systems increasingly rely on unstructured text analysis, our approach provides a robust foundation for next-generation enterprise clustering and recommendation systems.

This work represents a significant step toward more intelligent, semantically-aware business intelligence systems that can effectively navigate the complexities of modern global commerce.

REFERENCES

[1] Hsinchun Chen, Roger HL Chiang, and Veda C Storey. 2012. Business intelligence and analytics: from big data to big impact. *MIS quarterly* (2012), 1165–1188.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic BERT sentence embedding. *arXiv preprint arXiv:2007.01852* (2020).

[5] Christopher D Manning, Prabhakar Raghavan, and Hinrich Sch"utze. 2008. *Introduction to information retrieval*. Cambridge university press.

[6] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. http://arxiv.org/abs/1908.10084

[7] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 4512–4525.

[8] Stephen Robertson. 2004. Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation* 60, 5 (2004), 503–520.

[9] Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management* 24, 5 (1988), 513–523.

[10] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, Vol. 33. 5776–5788.