Similarity versus Supervision: Best Approaches for HS Code Prediction

Sédrick Stassin, Otmane Amel, Sidi Ahmed Mahmoudi and Xavier Siebert

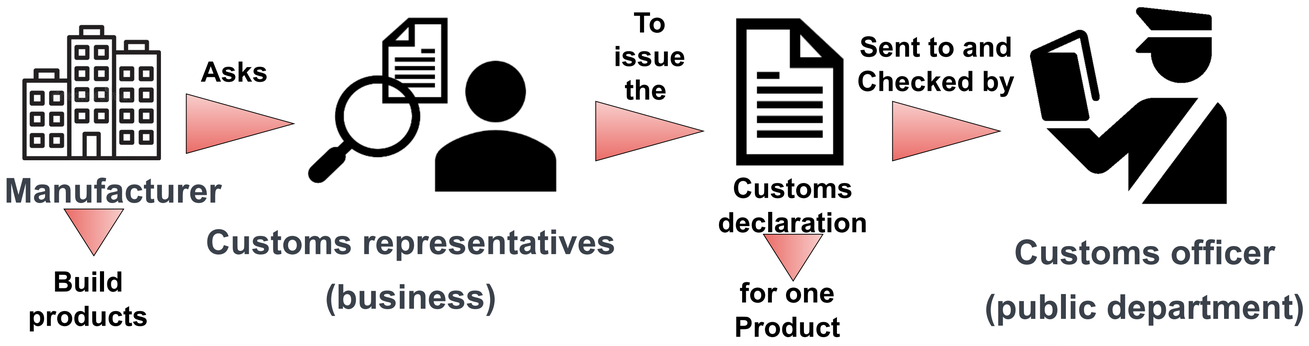
UMONS Université de Mons

inforTech INSTITUT DE RECHERCHE EN TECHNOLOGIES DE L'INFORMATION ET SCIENCES DE L'INFORMATIQUE DE L'UMONS

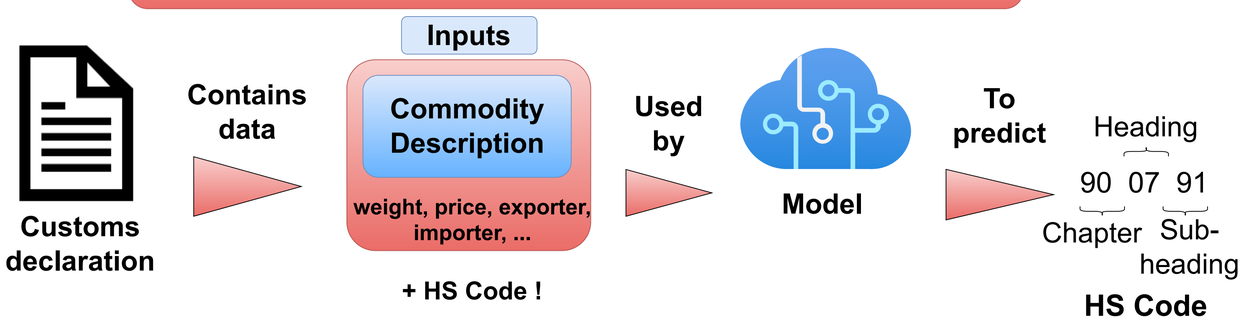
POLYTECH MONS

e-Origin

00 - THE CONTEXT

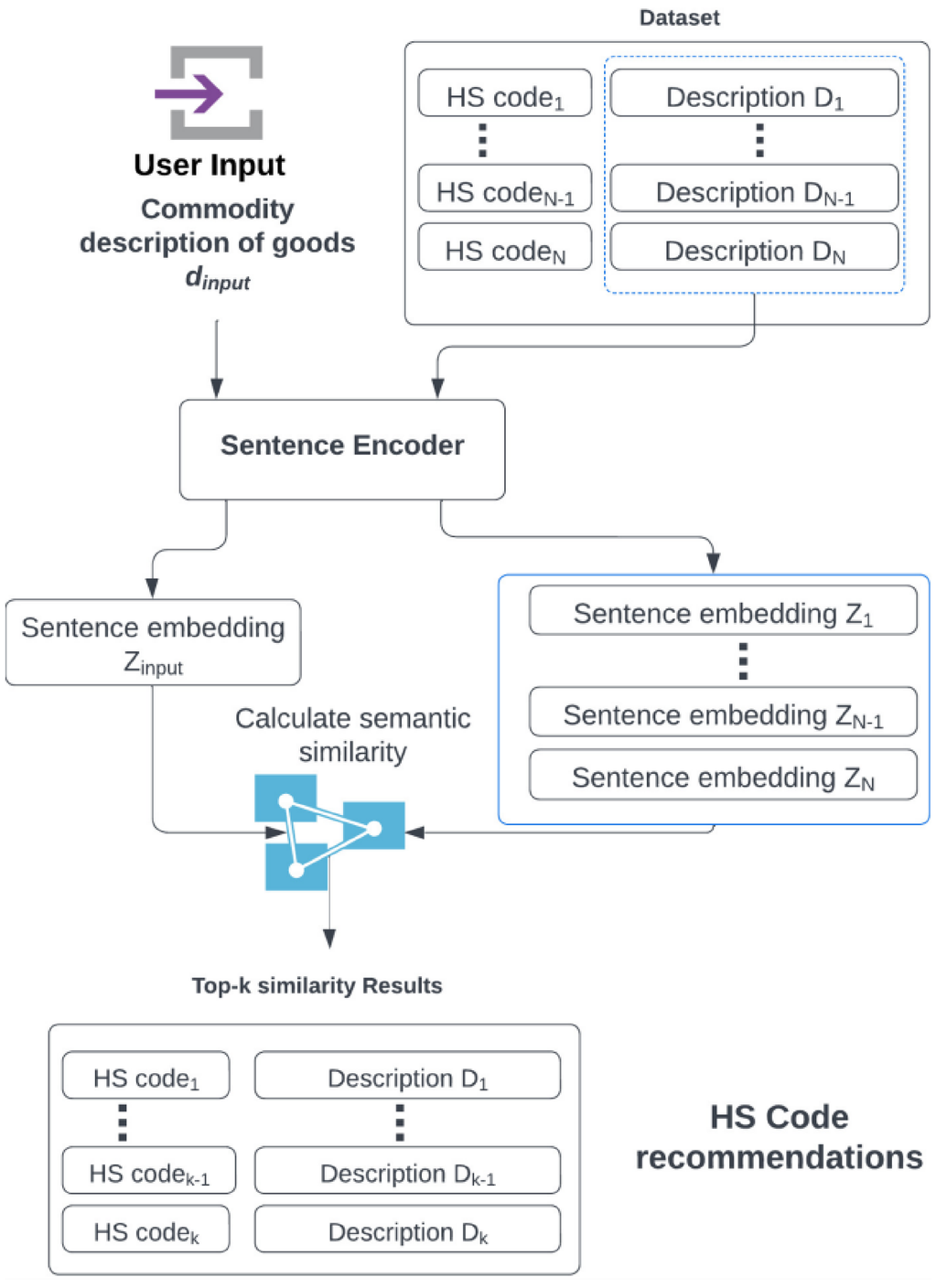


01 - THE TASK

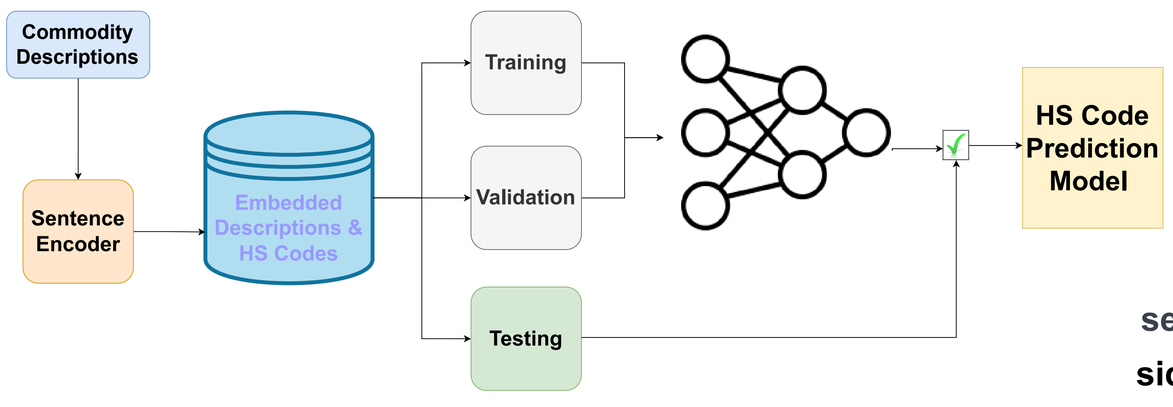


02 - THE APPROACHES

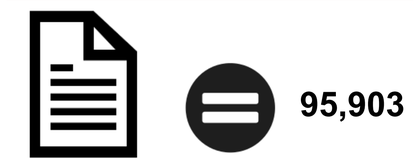
02.1 - SIMILARITY-BASED APPROACH



02.2 - SUPERVISION-BASED APPROACH



03 - DATASETS



04 - RESULTS

04.1 - SIMILARITY-BASED

# of digits	Top-k k=1	Top-k k=3	Top-k k=5	Model	Metric
6	**0,647**	0,900	0,956	MiniLM[3]	Chebyshev
	0,644	**0,910**	**0,961**	MiniLM[4]	Euclidean
	0,644	0,908	**0,961**	MiniLM[4]	Minkowski
8	**0,621**	0,878	0,950	DistilUSE[5]	Manhattan
	0,620	**0,889**	**0,952**	MiniLM[4]	Euclidean
	0,620	0,887	0,951	MiniLM[4]	Minkowski
10	**0,621**	0,883	0,947	DistilRoberta[6]	Manhattan
	0,618	0,880	0,947	MiniLM[7]	Manhattan
	0,617	**0,890**	**0,948**	MPNet[8]	Cosine

Table 1: Top-1, top-3, and top-5 accuracy of the semantic similarity-based approach with respect to the sentence embedding model and distance metric used.

- **Top-1 vs. Top3: many HS codes can be linked to one description**
- **<u>Resilience</u> to HS code complexity (HS6 to HS10)**
- **<u>Recommendation for each HS code</u> available**

04.2 - SUPERVISION-BASED

# of digits	Min. data / class	# of classes	Top-k k=1	Top-k k=3	Top-k k=5	Model
6	600	13	0.838	0.930	0.933	MiniLM[3]
	400	34	0.575	0.649	0.718	
	50	209	0.302	0.319	0.330	
8	600	13	0.844	0.885	0.918	MiniLM[9]
	400	33	0.633	0.662	0.664	
	50	209	0.282	0.294	0.316	
10	600	13	0.895	0.926	0.928	MiniLM[3]
	400	32	0.629	0.703	0.710	
	50	209	0.336	0.353	0.364	

Table 2: Top-1, top-3, and top-5 accuracy of supervised models with respect to the minimum number of samples per class.

- **<u>Performance drops with fewer min. samples per class</u>**
- **30% top-1 acc. for 209 classes ----- Similarity : 60% top-1 for 1,196 classes**
- **Added difficulty : Handling out-of-distribution classes**

06 - CONCLUSION

<u>Similarity-based approach</u> : MPNet achieves 89% and 94.8% top-3 and top-5 acc. across 1,196 classes for HS10 prediction

<u>Supervised-based approach</u> : worse performance + not scaling well with classes + not handling out-of-distribution classes

07 - CONTACT DETAILS

sedrick.stassin@umons.ac.be
otmane.amel@umons.ac.be
sidi.mahmoudi@umons.ac.be
xavier.siebert@umons.ac.be

Article QR code