We develop a method to train a document embedding model with an unlabeled dataset and low resources. Using teacher-student training, we distill SBERT's capacity to capture text structure and Paragraph Vector's ability to encode extended context into the resulting embedding model. We test our method on Longformer, a Transformer model with sparse attention that can process up to 4096 tokens. We explore several loss functions to enforce the distillation of knowledge from the two teachers (SBERT and Paragraph Vector) to our student model (Longformer). Throughout experimentation, we show that despite SBERT's short maximum context, its distillation is more critical to the student's performance. However, as we also demonstrate, the student model can benefit from both teachers. Our method improves Longformer's performance on eight downstream tasks, including citation prediction, plagiarism detection, and similarity search. Our method shows exceptional performance with few finetuning data available, where the trained student model outperforms both teacher models. By showing consistent performance of differently configured student models, we demonstrate our method's robustness to various changes and suggest areas for future work.

V této práci představujeme metodu strojového učení modelů emedující dokumenty, která není náročná na výpočetní zdroje ani nevyžaduje anotovaná trénovací data. S přístupem učitele a studenta, distilujeme kapacitu SBERTa zaznamenat strukturu textu a schopnost Paragraph Vektoru zpracovat dlouhé dokumenty do našeho výsledného embedovacího modelu. Naší metodu testujeme na Longformeru, Transformeru s řídkou attention vrstvou, který je schopný zpracovat dokumenty dlouhé až 4096 tokenů. Prozkoumáme několik ztrátových funkcí, které nutí studenta (Longformera) napodobovat výstupy obou učitelů (SBERTa a Paragraph Vektoru). Při experimentaci ukazujeme, že i přes omezený kontext SBERTa, je distilace jeho výstupů pro výkon studenta zásadnější. Nicméně, také ukazujeme, že student získává prospěch z obou učitelů. Naše metoda dokáže vylepšit výsledek Longformera na osmi úlohách, které zahrnují predikci citace, detekci plagiarismu i vyhledávání na základě podobnosti dokumentů. Naše metoda se navíc ukazuje jako obzvláště účinná v situacích s málo dotrénovávacími daty, kde námi natrénovaný student překoná i oba učitele. Podobným výkonem odlišně natrénovaných studentů ukazujeme, že naše metoda je robustní vůči různým změnám, kde dále navrhujeme možné oblasti budoucího výzkumu.