<table>
<tr><td rowspan="2"><b>cases</b></td><td><b>doc_1</b></td><td><b>doc_2</b></td><td><b>decision</b></td><td><b>id</b></td></tr>
<tr><td>
<table>
<tr><td><b>authors</b></td><td>- Song, X.
- Zhang, Z.</td></tr>
<tr><td><b>title</b></td><td>An exploratory study on utilising the web of linked data for product data mining</td></tr>
<tr><td><b>publication_date</b></td><td>2023-01-01 00:00:00</td></tr>
<tr><td><b>source</b></td><td>SupportedSources.CORE</td></tr>
<tr><td><b>journal</b></td><td></td></tr>
<tr><td><b>volume</b></td><td></td></tr>
<tr><td><b>doi</b></td><td>10.1007/s42979-022-01415-3</td></tr>
<tr><td><b>urls</b></td><td>- https://core.ac.uk/download/541480381.pdf</td></tr>
<tr><td><b>id</b></td><td>id1123915018276356290</td></tr>
<tr><td><b>abstract</b></td><td>The Linked Open Data practice has led to a significant growth of structured data on the Web. While this has created an unprecedented opportunity for research in the field of Natural Language Processing, there is a lack of systematic studies on how such data can be used to support downstream NLP tasks. This work focuses on the e-commerce domain and explores how we can use such structured data to create language resources for product data mining tasks. To do so, we process billions of structured data points in the form of RDF n-quads, to create multi-million words of product-related corpora that are later used in three different ways for creating language resources: training word-embedding models, continued pre-training of BERT-like language models, and training machine translation models that are used as a proxy to generate product-related keywords. These language resources are then evaluated in three downstream tasks, product classification, linking, and fake review detection using an extensive set of benchmarks. Our results show word embeddings to be the most reliable and consistent method to improve the accuracy on all tasks (with up to 6.9% points in macro-average F1 on some datasets). Contrary to some earlier studies that suggest a rather simple but effective approach such as building domain-specific language models by pre-training using in-domain corpora, our work serves a lesson that adapting these methods to new domains may not be as easy as it seems. We further analyse our datasets and reflect on how our findings can inform future research and practice</td></tr>
<tr><td><b>versions</b></td><td></td></tr>
</table>
</td><td>
<table>
<tr><td><b>authors</b></td><td>- Ziqi Zhang
- Xingyi Song</td></tr>
<tr><td><b>title</b></td><td>An Exploratory Study on Utilising the Web of Linked Data for Product Data Mining</td></tr>
<tr><td><b>publication_date</b></td><td>2022-06-24 00:00:00</td></tr>
<tr><td><b>source</b></td><td>SupportedSources.INTERNET_ARCHIVE</td></tr>
<tr><td><b>journal</b></td><td></td></tr>
<tr><td><b>volume</b></td><td></td></tr>
<tr><td><b>doi</b></td><td></td></tr>
<tr><td><b>urls</b></td><td>- https://web.archive.org/web/20220712072251/https://arxiv.org/pdf/2109.01411v4.pdf</td></tr>
<tr><td><b>id</b></td><td>id-2011699390288054886</td></tr>
<tr><td><b>abstract</b></td><td>The Linked Open Data practice has led to a significant growth of structured data on the Web in the last decade. Such structured data describe real-world entities in a machine-readable way, and have created an unprecedented opportunity for research in the field of Natural Language Processing. However, there is a lack of studies on how such data can be used, for what kind of tasks, and to what extent they can be useful for these tasks. This work focuses on the e-commerce domain to explore methods of utilising such structured data to create language resources that may be used for product classification and linking. We process billions of structured data points in the form of RDF n-quads, to create multi-million words of product-related corpora that are later used in three different ways for creating of language resources: training word embedding models, continued pre-training of BERT-like language models, and training Machine Translation models that are used as a proxy to generate product-related keywords. Our evaluation on an extensive set of benchmarks shows word embeddings to be the most reliable and consistent method to improve the accuracy on both tasks (with up to 6.9 percentage points in macro-average F1 on some datasets). The other two methods however, are not as useful. Our analysis shows that this could be due to a number of reasons, including the biased domain representation in the structured data and lack of vocabulary coverage. We share our datasets and discuss how our lessons learned could be taken forward to inform future research in this direction.</td></tr>
<tr><td><b>versions</b></td><td></td></tr>
</table>
</td><td>DUPLICATES</td><td>104</td></tr>
</table>