---------- REVIEW 1 ----------
SUBMISSION: 44
TITLE: Efficiently identifying pseudo-nulls in heterogeneous text data
AUTHORS: Théo Bouganim, Helena Galhardas and Ioana Manolescu

---------- Overall evaluation ----------
SCORE: 2 (accept)
----- TEXT:
The paper presents the first results of ongoing work regarding the identification of disguised null values in textual data collections modelled as ConnectionLens graphs. The proposal concerns the use of Machine Learning techniques to overcome the high cost of an initial method.

Strong points:

- The problem and challenges are clearly stated giving background in the relational context and then how it is redefined for other data models.
- The state of the art is particularly revealing going synthetically on existing strategies, approaches to the problem of detecting disguised null values from different perspectives.
- The experimental evaluation of the state of the art using the detector FAHES is clever and strategic. It is an original way of evaluating the scope and limitations of existing techniques.

Weak points:
- Section 2.2. even if FAHES is part of the state of the art and integrates different methods I did not get why it makes part of the state of the art. At some point, I consider this first experimental result as a kind of baseline that lets authors put in perspective the method they propose. I would devote a separate section to this first experiment about the state of the art solutions.

Answer:
- Sortir 2.2 de la 2
- Décrire le jeu de données PubMed (+ figure) dans une nouvelle section 3, "Motivating example"
- Renommer la 5 Experimental Evaluation
- Mettre les jeux de données HATVP dans 5.1
- 5.2 Null detection through FAHES

- At some point I was kind of lost on the way the problem is defined for the case of graphs. It was until I went through paragraph 3 of Section 3 that I got the idea. I think that it was kind of late, maybe the introduction can insist on this.

Answer: on peut dire dans l'intro The context of our work is ConnectionLens --> The motivation of our work came from the ConnectionLens. Mettre graph en bold dans cette phrase.

- Besides the first experimental results which seem promising, I would suggest better exploit the fact that the work is associated with a real application domain insisting on examples and maybe a use case.

Answer: créé section spéciale "Motivating example"

---------- REVIEW 2 ----------
SUBMISSION: 44
TITLE: Efficiently identifying pseudo-nulls in heterogeneous text data
AUTHORS: Théo Bouganim, Helena Galhardas and Ioana Manolescu

---------- Overall evaluation ----------
SCORE: 2 (accept)
----- TEXT:
The paper addresses the problem of identifying disguised nulls in text data. As opposed to explicit null values, which have received much attention in databases, disguised nulls require dedicated detection techniques, that are hard to generalize. And detecting disguised nulls is important in data cleaning and query processing to deliver correct answers. This work is done in the context of the ConnectionLens, a system for integrating heterogeneous, independently authored textual data sources in a single graph.

The authors first show that current methods in relational databases focus on numerical data and are not appropriate for textual data, where disguised nulls appear a lot, e.g., in fact-checking and journalism applications.

Then, they propose two methods:
. The first method leverages ConnectionLens's entity extraction, by establishing an entity profile for the text attributes of interest for null detection, and considering any value deviating from this profile a disguised null. This method is accurate, but inefficient because of complex entity extraction.
. The second method relies on text embeddings and classification, while also leveraging entity extraction but on a small subset of the dataset. This method is shown to be accurate and much more efficient than the first method.

Finally, the authors validate their proposal with a (short) experimental section
that studies the performance-precision trade-offs on real-world datasets using ConnectionLens.

Overall, a very important problem with a nice problem analysis and promising results. Even though it is a short paper, I would have preferred a longer experimental section, with more various datasets.

!!! MAR en 2.1 parle d'un "previous example" qui n'existe pas.

Minor

such as the Social Security one; => such as the Social Security database;

(2) A second method, leveraging ConnectionLens entity extrac-tion, is to (manually) …
Not sure what is the first method.

Answer
Section 1, outline:
- (2) devient: "(1) We show that CL entity extraction can be leveraged to..."

- (3) devint: "(2) To address this shortcoming, we devise a novel method..."
- On sépare la partie "We demonstrate..." et on y rajoute le (1) pour devenir un (3) We perform a set of experiments on the state-of-the art (FAHES). Our experiments show that..."

Revoir la dernière phrase de la 1.

Section 6 on related work is very short and should be merged with Section 2 (which you could call related work).

Answer: move current 6 content à la fin de la 2 - voir ce qu'il en reste (il peut y avoir des doublons).

---------- REVIEW 3 ----------
SUBMISSION: 44
TITLE: Efficiently identifying pseudo-nulls in heterogeneous text data
AUTHORS: Théo Bouganim, Helena Galhardas and Ioana Manolescu

---------- Overall evaluation ----------
SCORE: 1 (weak accept)
----- TEXT:
The paper present algorithms to identify disguised null values in text data. While this problem has been studied in structure data, there has been less attention to graph data with long text values. The authors propose two techniques to tackle this problem, with an emphasis on the performance.

The proposed optimization based on using embeddings is simple but reasonable and very effective. However, it relies on the assumption that original solution (based on entity profile) has an always accurate entity extractor. This may be challenging in real setting, but there seems to be two more important issues here
- the embedding based solution is measured wrt the original method treated as ground truth, not against the real ground truth
Answer:
- we do consider entity extraction as the ground truth.
- not much else we can do with text data and at these scales
- it's sometimes wrong but not much
- we analyze *sets* of strings thus some errors are probably absorbed (do not impact the global analysis)

- the qualitative analysis done for [21] in sec 2 is not reported for these proposed methods

I would suggest the authors to compare the baseline and the proposed methods in the same conditions. From the discussion in sec 2, it seems the baseline is doing ok, but has limitation. How good is it? can we get a quantitative evaluation for it? we should then measure the proposed method against the same ground truth and with the same method. Then it will be more interesting to see the benefit of the embedding optimization in terms of time saving with little loss in quality.

Answer:
Venn diagram and one example for each area in the diagram (example from the paper or not)

Discussion section (or paragraph) après avoir présenté notre méthode

The paper is already interesting and has some nice ideas, with some clarifications on these evaluations aspects it would be much stronger in making the case for this line of research.

À faire:
- Changer le diagramme et donner une taxonomie précise des DMVs qu'on détecte.
- Refaire table 2 et rajouter la comparaison avec sentenceBert (considérer embedding time)
- Rajouter dans l'intégration une explication de pourquoi on a choisi tf-idf et pas sent