Ankita Biswas (ab8ky@virginia.edu)
DS5001
06 May 2022

An investigation of scientific journals exploring structure-property relationships of High-Entropy Alloys

Ankita Biswas[a], Joseph B Choi[b]
[a]Department of Materials Science and Engineering, University of Virginia
[b]School of Data Science, University of Virginia

1. Introduction

High entropy alloys (HEA) are a new breed of alloys which are different from conventional metallic alloys w.r.t the fact they contain at least five principal elements with concentrations distributed such that usually an equiatomic ratio is maintained[1]. The high mixing entropy of these new alloys is predicted overcome the enthalpies of the compound formation and leading to the formation of the simple disordered solid solution crystal structure of face-centered (**fcc)**, body-centered (**bcc)** and Hexagonal close-packed (**hcp)** [2]**.** In this study we have tried to implement various exploratory text analytics techniques to a corpus compiled from 15 seminal scientific articles from several well known scientific literature publishing organizations.

2. Information generation

The original idea was to build a larger corpora of articles of papers both from HEA and non-HEA related papers. But given the complexity of data pre-processing from scientific journals, the work was limited to only HEA related papers. Kononova et al.[3] previously did a similar data extraction work based on other inorganic chemical compounds using source files in the HTML format. However, in this study the data extraction from the pdf format as the older articles pertaining to the High-Entropy alloys (before 2000) are not found in the HTML format. Belval's *pdf2image* package in python was utilized to convert the pdf to image format. Then Lee's pytesseract 0.3.9 package in python was utilized to extract the textual information from the image file. However, the textual information thus extracted contained several unwanted elements such as text from the scientific figures and tables as well as equations. Thus an iterative process was formulated to reduce the noise in the text convert the same to a workable format.

The iterable steps included the following:

i. Removal of special characters and irrelevant sections (acknowledgements & references).
ii. Replaced multiple new lines into a single new line in between the lines.
iii. Combined new lines into paragraphs with the new line as a delimiter and hyphens at the end of the sentence where words were cutoff, were dealt with.
iv. Captions of figures and tables were removed.
v. Other extracted textual information from figures, tables and equations were removed.
vi. Paragraphs that were split due to cleaning of noise arising from the above mentioned factors were combined together.
vii. Some manual removal of special characters and numerical characters from within the paragraphs were also necessary step for several papers.
viii. Finally the paragraphs, sentences and words were combined into *token* tables for each paper according to appropriate *OHCO* levels. These *token* tables were further combined leading to the final corpus.

3. Results and Discussions

I. TFIDF

The “Term Frequency-Inverse Document Frequency” (TF-IDF) values for the entire corpus was calculated.This is a way of extracting information about the most significant words in the corpus overall. The top 20 words by TF-IDF mean score in the overall corpus were first computed as shown below.

term_str	mean_tfidf	max_pos
constants	0.040556	NNS
elastic	0.032070	JJ
elasticity	0.037104	NN
magnetic	0.041073	JJ
milling	0.062640	NN
misfit	0.030958	NN
mn	0.028162	NNP
mno	0.049797	NNP
modulus	0.030140	NN
nbmotaw	0.028649	NNP
nico	0.027717	NNP
nicofemncr	0.038595	NNP
nm	0.039823	NN
rhea	0.044201	NNP
sigma	0.027422	NN
sss	0.028081	NNP
superconducting	0.028870	VBG
theory	0.029946	NN
variation	0.029776	NN
volumes	0.034498	NNS

Figure 1. TFIDF sorting of most significant terms (top 20)

The most significant word comes out to be ‘milling’ which is known to be one of the most important manufacturing techniques of this class of alloys. The next most significant word came out to be ‘mno’ or MnO (manganese oxide), manganese one of the most important elements of the CoFeNiMnCr HEA family which plays a vital role in the phase stability, mechanical properties and oxidation behavior. Some other significant words being ‘*constants*’ presumably indicating to the lattice constant, the structural imformation that influences different mechanical properties, ‘*magnetic*’ a functional property observed in many of these compounds, ‘*elasticity*’ a very important mechanical property that influences

the hardness/strength of these alloys which is the very property to look out for in these alloys. This might already start giving a primary idea to the uninitiated in the domain about the HEA alloy system.

Similarity between documents was explored through dendogram clustering based on the tfidf scores.

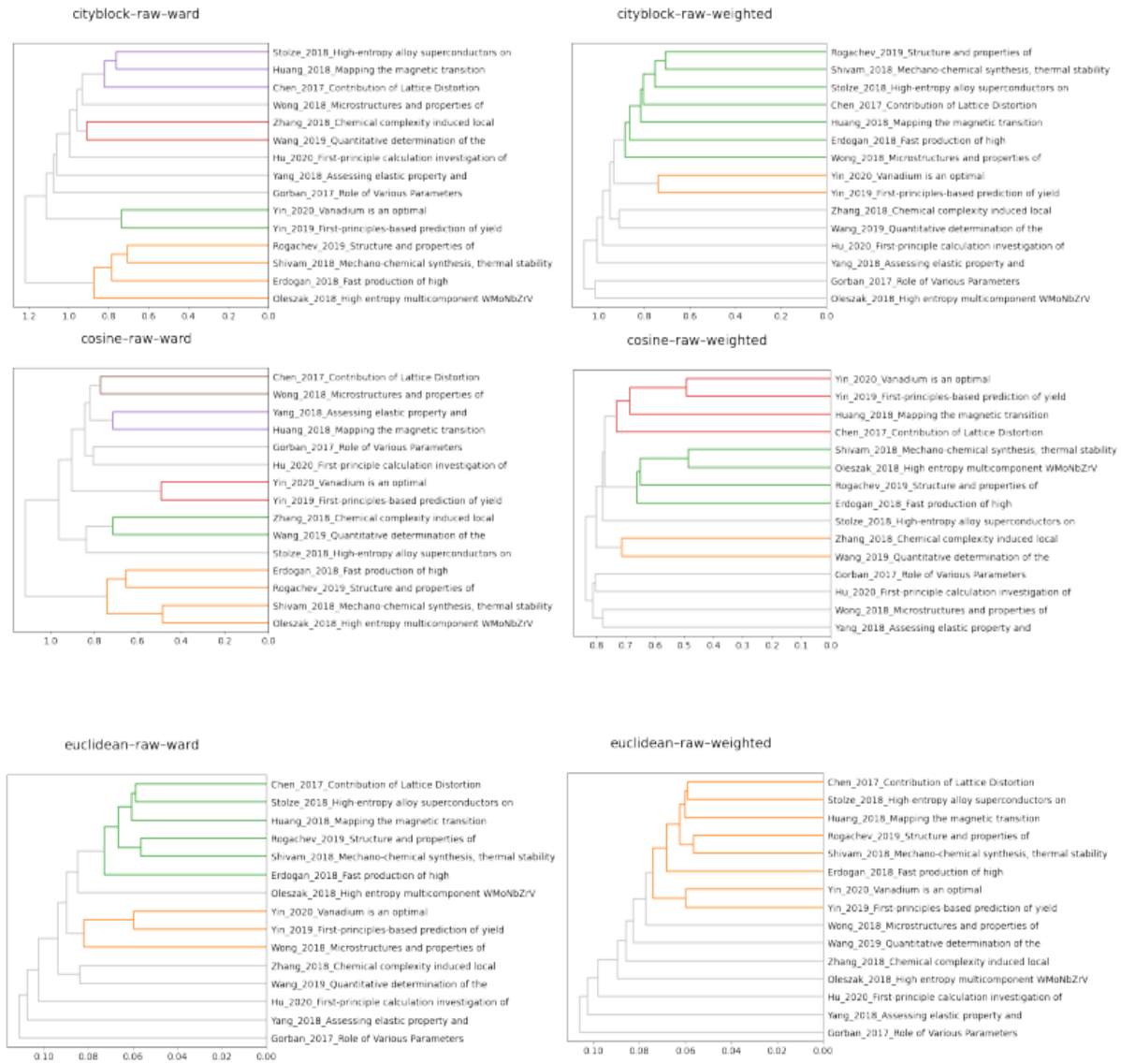


Figure 2. Different similarity measures explored in the study w.r.t tfidf. (Figures generated by J. Choi)

Clearly the euclidean distance measure gives a better idea of the distinction between the clusters due to sensitivity to the document length. Interesting thing to note was the clustering of the Yin-Curtin computational papers together. This showed some reference made to the methodologies too.

II. Principal Component Analysis (PCA)

PCA was implemented to explore the most informative dimensions in the reduced feature space. The result is shown as follows:

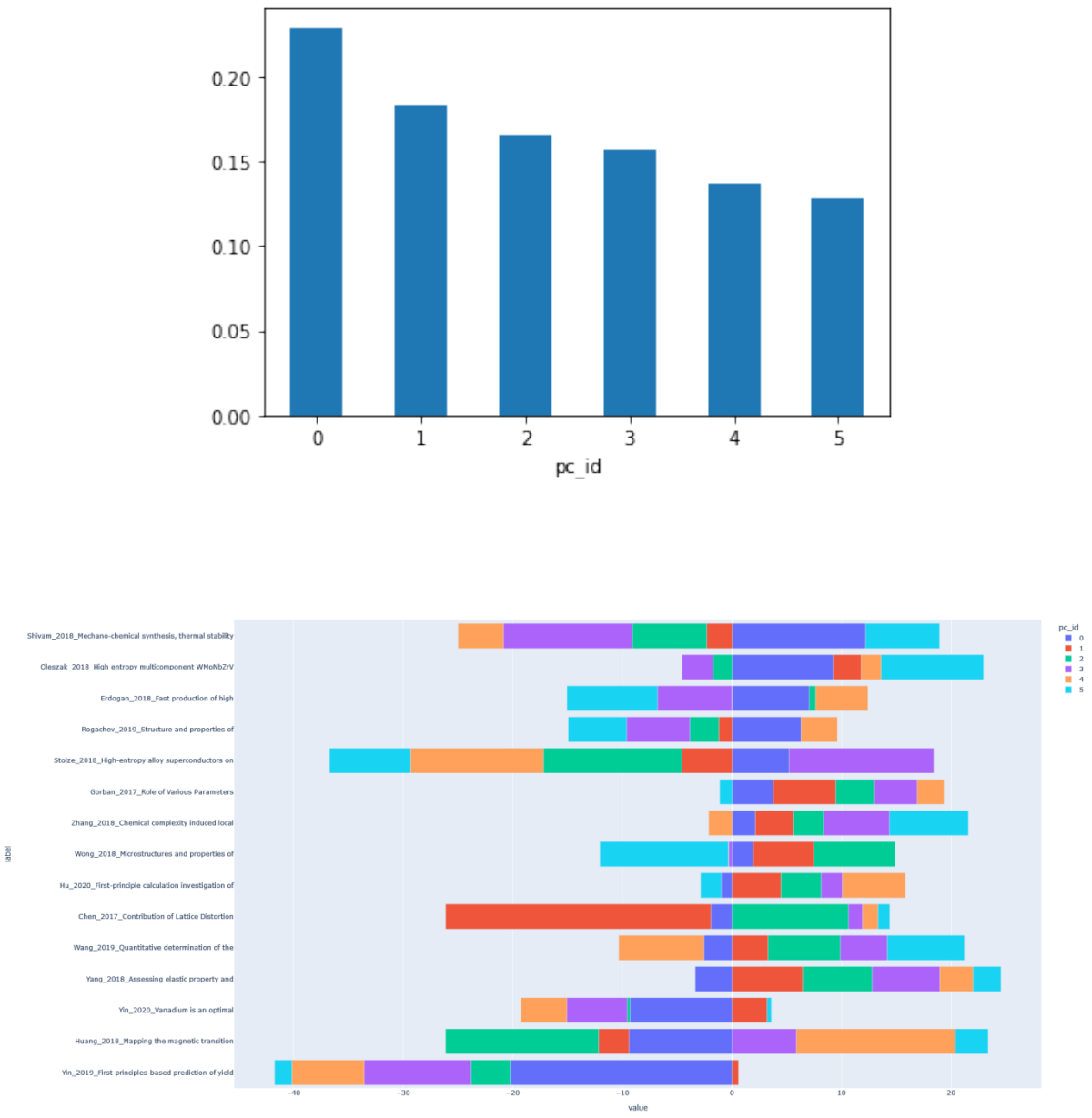


Figure 3. Principle component 0 explains maximum variance. Clearly the new feature space obtained from PCA shows a clear distinction between the sources of information, however, the interpretation is debatable at most. 2nd part of figure generated by (J. Choi).

The interpretation of why one component (component 1) draws more information from one particular article (Chen 2017) could not be made very clear. This gives an idea of the lack of

interpretability of PCA.

III. Topic Modeling

For topic modeling the *Latent Dirichlet Allocation* (LDA) method was implemented. Results are given below for the first 10 topics computed all over the corpus by the document-weight (theta) sum.

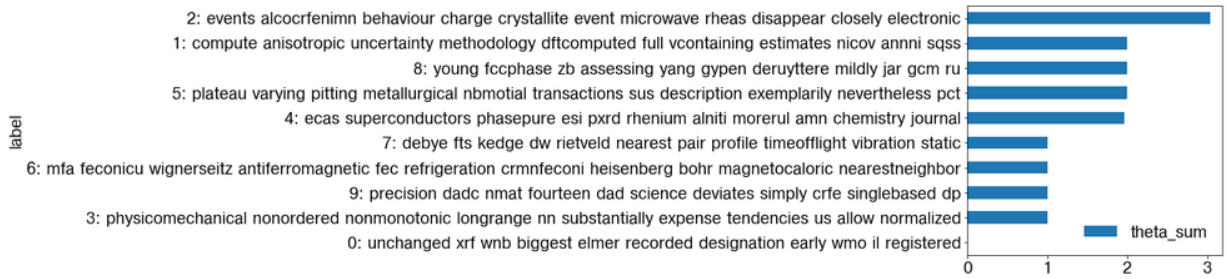


Figure 4. LDA applied to the BOW generated for the document.

The first topic that catches the eye easily is topic 8. The terms ‘fcc phase’ and ‘young’ corresponding to the Youngs Modulus (YM) of a material are very much related. Ductility is the property of a material to be able to pulled into wires hence characteristic high plastic deformation. Usually FCC crystals show this behaviour gold, copper, silver etc. Now you are asking a relation between young’s Modulus and ductility. See YM is the slope of a stress strain plot. There is a region where the Stress is proportional to the strain and the relationship is linear (elastic region) You can calculate YM in that region to have a precise value . Once the lenearity is dissolved (when plastic deformation starts) estimations of YM become inaccurate and then you get point slope of a curve at various points. Now high YM means the elasticity is high and ductile materials are not elastic rather they are plastic because they are predominantly soft. High YM means your material under observation is able to resist well (tool grade steel, composites,rocks, wood etc.) the external loading and hence the fcc materials in short are highly ductile materials and show very little YM.

High ductility → low YM , its not a direct relationship but rather the science of it and interpretation. The topic 8 thus justifies the above interpretation. Another interesting topic would be topic no 6. With topics like ‘magnetocaloric’ and FeCoNiCu system of compounds mentioned it kind of references the magnetocaloric functional properties of the above class of compounds. This would be highly interesting to follow up with further analysis on a broader spectrum of topics.

IV. Word Embeddings (word2vec)

This was implemented to explore the relationships between words existing in the corpus. The results are shown below:

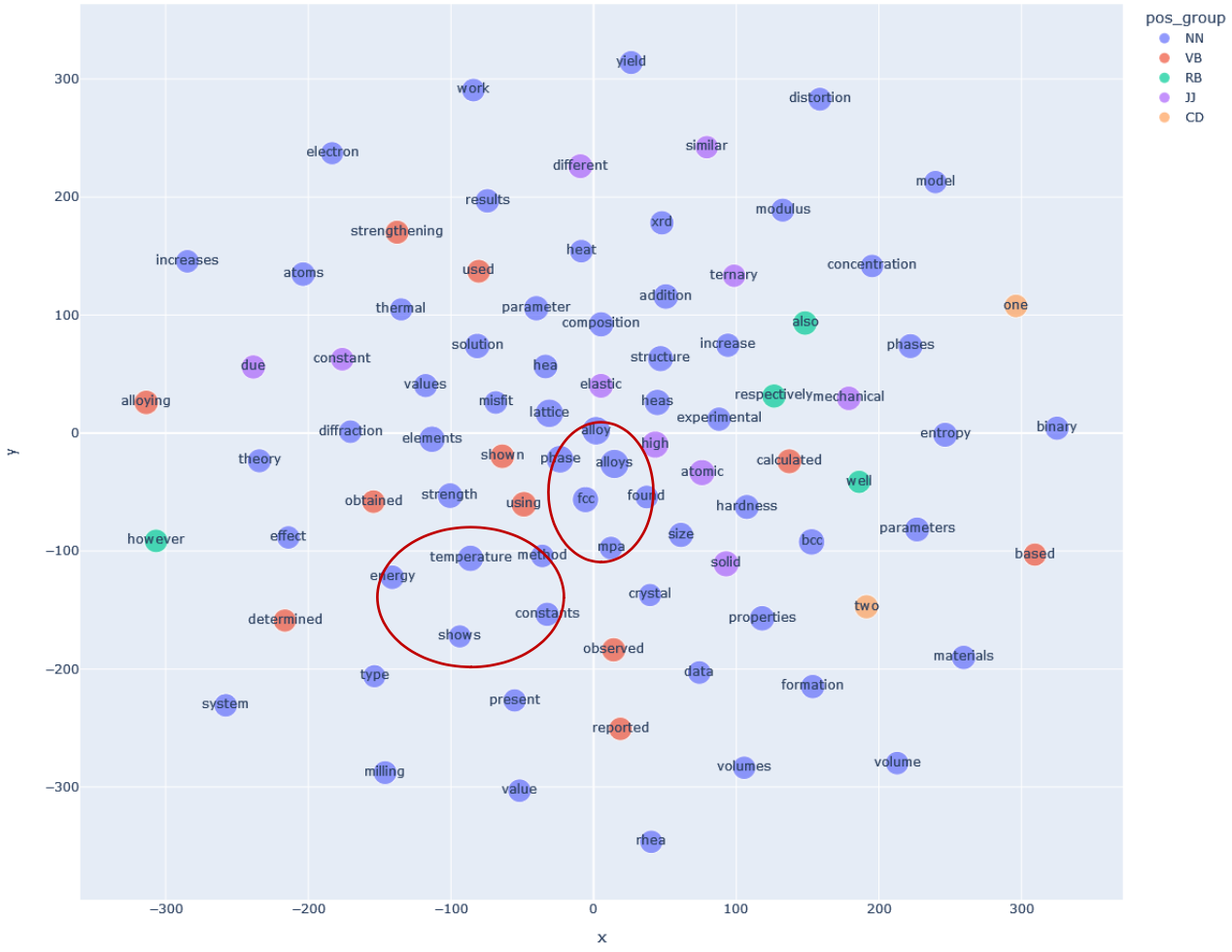


Figure 5 t-SNE plot generated after application of word2vec model.

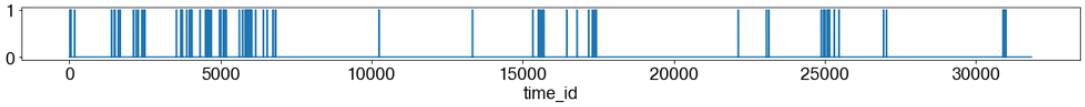
An interesting cluster of words appear around (-40,20): [lattice, alloy, phase, alloys, fcc] -> gives structural information mostly. Another interesting cluster appears around (-160, -20): [method, temperature, energy, constants, shows] -> gives mainly manufacturing parameter information. Another highly interesting trend w.r.t HEAs captured by text analytics.

V. Token-Time matrix

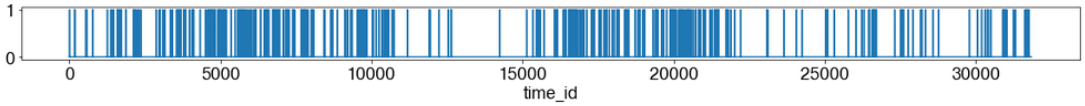
An attempt was made to visualize the dispersion plots of words. Results are as shown below:

```
In [43]: TTM['lattice'].plot(**cfg);
```

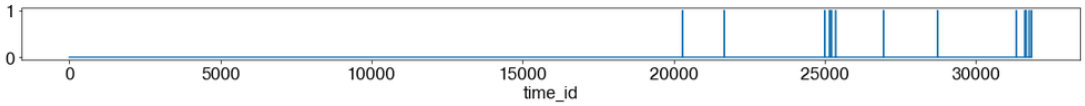

```
In [44]: TTM['entropy'].plot(**cfg);
```




```
In [45]: TTM['alloy'].plot(**cfg);
```

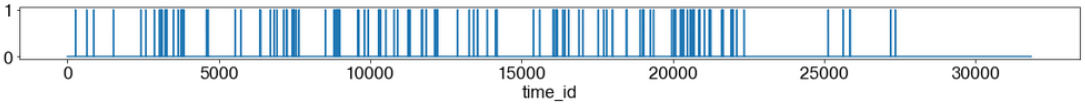



```
In [46]: TTM['refractory'].plot(**cfg);
```

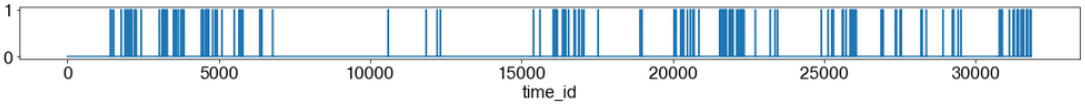



```
In [47]: TTM['hardness'].plot(**cfg);
```


```
In [48]: TTM['fcc'].plot(**cfg);
```




```
In [49]: TTM['bcc'].plot(**cfg);
```




```
In [50]: TTM['hcp'].plot(**cfg);
```

This shows some interesting trends w.r.t higher research on ‘fcc’ alloys than other structures. Also significantly high had been the research on ‘lattice’ which includes not only the lattice structure of the compound but also the lattice parameter information. This can be a very important trend to motivate and channelize more resources towards studies of other lattice structures and properties than what is shown above.

VI. Conclusion

Some highly interesting trends and information came up for the High-entropy alloy based corpus. It was realized the analysis could have benefitted more if some non-HEA related papers were included in the corpus. However, due to time constraint and increased complexity of data pre-processing of scientific literature, the work was limited to only HEA related papers. Even though sentiment analysis could not be performed, the already existing

trends would easily help the uninitiated to get a very good idea about HEA alloy systems - an extremely important class of compounds in materials science well known for their hardness, strength and in general good refractory properties.

References

1. D. Oleszak, A.A-Dudka, T. Kulik, *'High entropy multicomponent WMoNbZrV alloy processed by mechanical alloying'*, Materials Letters 232 (2018) 160–162.
2. V. Shivam , J. Basu, Y. Shadangi, M.K. Singh, N.K. Mukhopadhyay, '*Mechano-chemical synthesis, thermal stability and phase evolution inAlCoCrFeNiMn high entropy alloy*', Journal of Alloys and Compounds 757 (2018) 87e97.
3. O. Kononova, H. Huo, T. He, Z. Rong, T. Botari, W. Sun, V. Tshitoyan, G. Ceder, '*Text-mined dataset of inorganic materials synthesis recipes*', Sci Data 6, 203 (2019).