Concatenated Power Means Sentence Embedding in Unsupervised Scientific Paper Clustering

Abstract - Academic Paper Retrieval systems are widely used in academic institutions to store and categorize Scientific papers and articles. There is an extensive literature about clustering these stored papers into categories and finding connections between them using citation links, but these old methods do not account for the content of the papers. We propose and verify a new method for unsupervised clustering of papers by using Concatenated Power Means Sentence embeddings of abstracts using Natural Language Processing.

1 INTRODUCTION

Since "Networks of Scientific Papers" [1] was published in 1965, there have been many systems and research focused on clustering academic papers by considering links between them. Academic papers have been clustered using Co-Citation Analysis, Bibliographic Coupling, and Direct Citation relations [2] [3] [4], subject-based algorithmic classification of different granularities [5], by using non-parametric methods such as Adaptive Weights Clustering on their JEL classification tags [2], [6]. But all of these methods do not take the actual content of the papers into account for clustering and categorizing them.

With latest advancements in Natural language processing, several ways of clustering papers, documents, and texts are being explored e.g. clustering Biomedical publications using TF-IDF, Latent Semantic Indexing, Topic Modeling [7], clustering social media posts for Pharmacovigilance using word embeddings [8].

Word embeddings such as GloVe [9], Word2Vec [10], FastText [11] are the representations of words in an n-dimensional vector space. Sentence embeddings are dense vectors that summarize different properties of a sentence (e.g. its meaning), thereby extending the very popular concept of word embeddings to the sentence level.

We propose and verify unsupervised clustering of academic papers for exploratory data analysis by using the Concatenated Power Means [12] (pmeans) and a simple Centroid of their abstracts as the clustering keys. We use word embeddings (Glove, Word2Vec, and FastText) of words in the abstract and create a single sentence embedding from them, which will be later used to cluster them using K Means, Mini Batch K means and Spectral Clustering algorithms. We will then analyse these methods by statistical and performance measures.

In this paper, we first explore the methodology used to create these embeddings and cluster the papers, followed by statistical and performance measures of clustering and finally a discussion on the viability of our approach.

2 METHODOLOGY

A. Dataset

We collected the dataset from https://core.ac.uk/services/dataset/, the website offers multiple datasets varying in the year of publication and total size. We chose the metadata dataset published in 2013 as it suited our need of having just the abstract, as we did not have any use of the body for clustering.

B. Sentence Embeddings

The abstracts were lemmatized, stop words were removed and converted to arrays of word embeddings For every word embedding type. We used GloVe 50d, Word2Vec 100d, and FastText 300d word embeddings.

Concatenated Power Means sentence embedding and centroid sentence embeddings were created for every word embedding type. The Power Means method generalizes the average word embeddings by retrieving many well-known means such as the arithmetic mean (p=1), the geometric mean (p=0), and the harmonic mean (p=−1). When $p=\pm\infty$, the power

means specializes in the minimum (p=−∞) and maximum (p= +∞) of the sequence.

$$\left(\frac{x_1^p + \cdots + x_n^p}{n}\right)^{1/p} ; \; p \in R \cup \{\pm\infty\}$$

Here x is word embedding and n is the number of words.

We used p=1, +∞, −∞, 2 and 4.

We chose pmeans because of its established accuracy in multiple downstream tasks and low computational power requirement [12] [13].

centroid Sentence embedding is power means with p=1.

$$\left(\frac{x_1 + \cdots + x_n}{n}\right)$$

We compared pmeans with the simplest sentence embedding algorithm i.e. centroid.We did not compare it with SIF-Sentence Embeddings [14] since the number of words repeated in paper abstracts is very less.

We used Kmeans, MiniBatch Kmeans, and spectral clustering algorithms.

Selection of number of clusters (K) using Zipf's law We calculated silhouette score for range K = 2, 6, 10, 14, …., 3000 (skipping every 4 digits) with 30000 papers for Kmeans using Pmeans with GloVe. We found out that for every multiple of the square root of 30000 (i.e. 173.2 ≈ 173) the Silhouette score did not increase substantially after 346 (i.e. twice the square root). The Silhouette score was 0.1 to 0.8 for K = 2 to 18. But these did not give us enough clusters from an application point of view.

We trained our clustering model in the following combinations - Algorithm: K Means, Mini Batch K Means, Spectral. Word Embedding: GloVe 50d, Word2Vec 100d, FastText 300d. Sentence Embeddings: pmeans, centroid. Cluster and number of papers pairs: (50, 3000), (250, 15000), (350, 30000). 250 and 350 are approx twice the square root of 15000 and 30000 respectively So there are a total of 54 models.

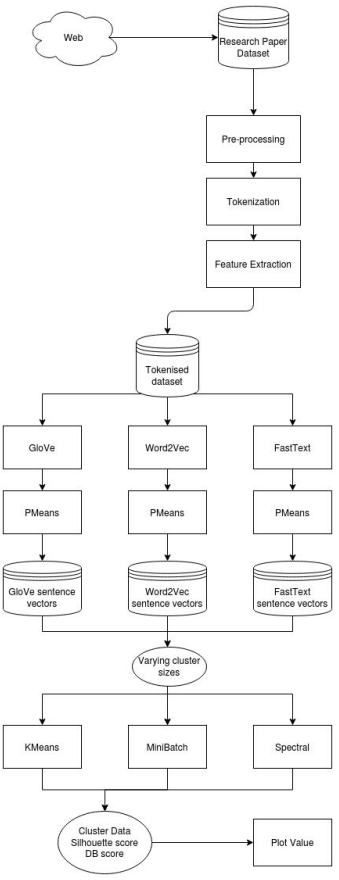

Fig.1. Research Pipeline

C. Clustering

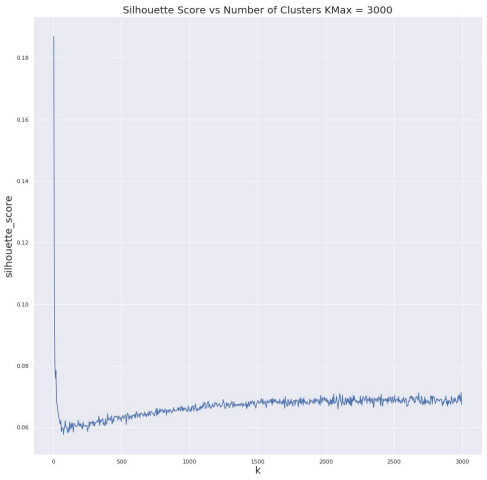

Fig. 2. Silhouette Score Vs Number of Clusters

We compared the Silhouette score and Davies Bouldin Score of every model.

3 RESULTS

We will compare the Silhouette score and Davis Boulding score for models trained on 30000 papers for K Means models as that represents the real-world data.

The results delivered by Mini batch K means and Spectral were very poor as compared to Kmeans (almost half of Kmeans in every case). Also in spectral clustering, we have to train the model every time we have to predict the label for a test or new paper, which is very time-consuming. So in results, we mainly focus on K means.

A. Clustering Scores

Sentence Embedding Algorithm	Word Embedding	Silhouette Score	Davis Bouldin Score
Centroid	GloVe	0.0814	2.4284
Centroid	word2Vec	0.0740	2.5006
Centroid	FastText	0.0657	2.7902
Pmeans	GloVe	0.0631	3.5175
Pmeans	word2Vec	0.0621	3.5023
Pmeans	FastText	0.0574	3.8057

Table 1. Model comparison for 30000 papers and 350 Clusters for Kmeans

Sentence Embedding Algorithm	Word Embedding	Silhouette Score	Davis Bouldin Score
Centroid	GloVe	0.0816	2.3878
Centroid	word2Vec	0.0776	2.4808
Centroid	FastText	0.0674	2.7141
Pmeans	GloVe	0.0635	3.3782
Pmeans	word2Vec	0.0617	3.3748
Pmeans	FastText	0.0595	3.7380

Table 2. Model comparison for 15000 papers and 250 Clusters for Kmeans

Sentence Embedding Algorithm	Word Embedding	Silhouette Score	Davis Bouldin Score
Centroid	GloVe	0.0871	2.2623
Centroid	word2Vec	0.0770	2.3245
Centroid	FastText	0.0735	2.5742
Pmeans	GloVe	0.0643	3.1729
Pmeans	word2Vec	0.0604	3.1012
Pmeans	FastText	0.0331	3.3654

Table 3. Model comparison for 3000 papers and 50 Clusters for Kmeans

We can see that the Silhouette score calculated for the clusters is very low. An acceptable silhouette score is considered to be above 0.25 and a good score is considered above 0.6. None of the models trained by us scored more than 0.1 silhouette score.

B. Cluster Distribution

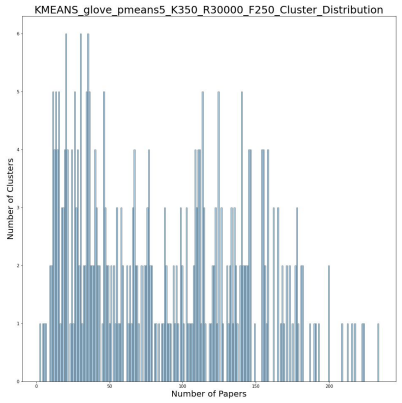

Fig 3. Cluster Distribution of 30000 papers

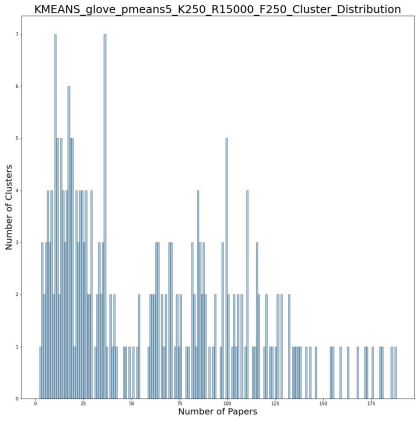

Fig 4. Cluster Distribution of 15000 papers

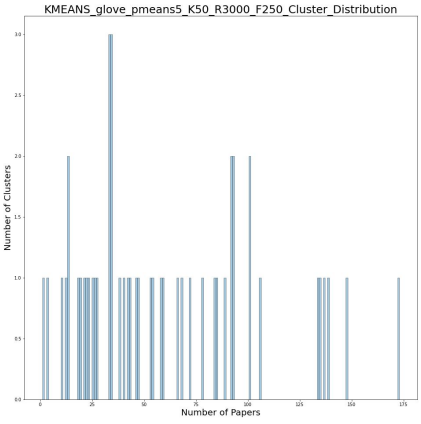

Fig 5. Cluster Distribution of 3000 papers

We can see in Fig. 3, 4, 5 that in the case of all numbers of papers, the clusters follow a similar distribution. But as the number of papers increases the number of clusters having less number of papers also increases.

C Similarity

Word Emb and Sentence Emb.	Paper Titles (Refer Appendix for full Abstracts)	Cosine Distance from Test Paper
Test Paper Title	Intracellular mechanisms underlying the nicotinic enhancement of LTP in the rat dentate gyrus	0 (Test Paper title)
GloVe + Centroid	The novel Syk inhibitor R406 reveals mechanistic differences in the initiation of GPVI and CLEC-2 signaling in platelets	0.0486
	an investigation into the role of neurotransmitter receptors in the function of human immune cells	0.0599
GloVe + Pmeans	Phenylephrine preconditioning of isolated ventricular myocytes involves modulation of KATP channels through activation of survival kinases	0.0253
	The tyrosine phosphatase CD148 is an essential positive regulator of platelet activation and thrombosis	0.0254
Word2Vec + Centroid	GLP-1 and Muscarinic Receptor Mediated Activation of ERK1/2 in Pancreatic β-cells	0.0271
	Biochemical investigation of phosphodiesterase type IV post-translational modification, cellular localisation and interaction with associated binding proteins	0.0280
Word2Vec + Pmeans	A population of immature cerebellar parallel fibre synapses are insensitive to adenosine but are inhibited by hypoxia	0.0247
	Excitotoxic ATP and Glutamate Signalling during Central Nervous System Ischaemia	0.0251
Fastext + Centroid	Phenylephrine preconditioning of isolated ventricular myocytes involves modulation of KATP channels through activation of survival kinases	0.0429
	Opposing Changes in Phosphorylation of Specific Sites in Synapsin I During Ca2-Dependent Glutamate Release in Isolated Nerve Terminals	0.0447
Fastext +	Scanning peptide array analysis	0.0256

Pmeans	identify overlapping binding sites for the signaling scaffold proteins, beta-arestin and RACK1 in the cAMP-specific phosphodiesterase, PDE4D5	
	Novel areas of crosstalk between the cyclic AMP and PKC signalling pathways	0.0270

Table 4. Results of K means clustering, for Top 2 nearest research papers in paper's cluster

As we can see in Table 4, The clustered papers are in the same field of literature and the cosine distance between the closest research papers is low.

4 DISCUSSION

The Silhouette score of all the above models is less than 0.1, averaging about 0.06. This is very subpar. Acceptable Silhouette score should at least be 0.25. In some cases of our models, it was even negative. Even with the lower value of k (2 to 18), the score never goes above 0.1. Also after looking at the distributions, we can say that there are many small clusters as compared to large clusters. Ideally, there should be many average-sized clusters.

The clustered research papers are very similar, they belong to the same field of science. This posits a hypothesis that the similarity of related research papers is calculated correctly, but the nature of the dataset of research papers does not allow an acceptable clustering model to be trained.

From our clustering Experiments, it is apparent that the unsupervised clustering of research papers using Concatenated Power means and Centroid Sentence embedding is unsatisfactory and should not be used.

5 CONCLUSION

Older and more tried, researched clustering techniques such as Co-Citation Analysis, Bibliographic Coupling, and Direct Citation relations are better for unsupervised clustering of research papers as compared to Clustering using our proposed technique. Clustering using Natural Language Processing techniques such as keywords, LSI, LDA, topic modeling also give better statistical outcomes than our proposed method. Unsupervised Clustering using Concatenated Power Means and Centroids does not give sufficient statistical score to be deemed viable for unsupervised clustering of research papers. It is more viable for supervised clustering as it gives good, logical similarity scores for research papers in the same cluster, but the nature of the dataset makes it harder for unsupervised clustering to be statistically viable.

6 FUTURE WORK

For clustering and classifying research papers using sentence embeddings, neural network based sentence embedding models such as Universal Sentence Encoder [15] or Infersent [16] can be used. Concatenated Power means should be tested on a labeled dataset i.e. supervised clustering, to get accuracy, precision and recall scores to create comparable metric.

ACKNOWLEDGMENTS

We acknowledge the guidance, academic and computational resources provided by Dr. Prof. Y. V. Haribhakta, Department of Computer Engineering and Information Technology, College of Engineering, Pune.

REFERENCES

[1] D. J. de S. Price and D. J. de Solla Price, "Networks of Scientific Papers," *Science*, vol. 149, no. 3683. pp. 510–515, 1965, doi: 10.1126/science.149.3683.510.

[2] L. Šubelj, N. J. van Eck, and L. Waltman, "Clustering Scientific Publications Based on Citation Relations: A Systematic Comparison of Different Methods," *PLoS One*, vol. 11, no. 4, p. e0154404, Apr. 2016.

[3] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12. pp. 2389–2404, 2010, doi: 10.1002/asi.21419.

[4] R. Klavans and K. W. Boyack, "Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?," *Journal of the Association for Information Science and Technology*, vol. 68, no. 4. pp. 984–998, 2017, doi: 10.1002/asi.23734.

[5] P. Sjögårde and P. Ahlgren, "Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics," *Journal of Informetrics*, vol. 12, no. 1. pp. 133–152, 2018, doi: 10.1016/j.joi.2017.12.006.

[6] L. Adamyan, K. S. Efimov, C. Chen, and W. K. HHrdle, "Adaptive Weights Clustering of Research Papers," *SSRN Electronic Journal*. doi: 10.2139/ssrn.2997061.

[7] K. W. Boyack *et al.*, "Clustering More than Two Million

Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches," *PLoS ONE*, vol. 6, no. 3. p. e18029, 2011, doi: 10.1371/journal.pone.0018029.

[8] A. Nikfarjam, A. Sarker, K. O'Connor, R. Ginn, and G. Gonzalez, "Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features," *Journal of the American Medical Informatics Association*. 2015, doi: 10.1093/jamia/ocu041.

[9] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, doi: 10.3115/v1/d14-1162.

[10] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," *https://arxiv.org*, Sep. 2013, doi: arXiv:1301.3781.

[11] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. 2017, doi: 10.18653/v1/e17-2068.

[12] **Andreas Rücklé , Steffen Eger , Maxime Peyrard , Iryna Gurevych, "Concatenated Power Mean Word Embeddings as Universal Cross-Lingual Sentence Representations," *https://arxiv.org*, Department of Computer Science, Technische Universität Darmstadt, 12-Sep-2018.**

[13] Christian S. Perone, Roberto Silveira, Thomas S. Paula, "Evaluation of sentence embeddings in downstream and linguistic probing tasks," *ICLR 2017*, Jun. 2018, doi: arXiv:1806.06259.

[14] Y. L. Sanjeev Arora, "A Simple but Tough-to-Beat Baseline for Sentence Embeddings," *ICLR 2017*, Nov. 2016.

[15] Daniel Cer, Yinfei Yang, Sheng-yi Kong, et al., "Universal Sentence Encoder," *Google Research*.

[16] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017, doi: 10.18653/v1/d17-1070.