Corpus Statistics Empowered Document Classification

Subjects: Computer Science, Artificial Intelligence

Contributor: Farid Uddin, Yibo Chen, Zuping Zhang, Xin Huang

In natural language processing (NLP), document classification is an important task that relies on the proper thematic representation of the documents. Gaussian mixture-based clustering is widespread for capturing rich thematic semantics but ignores emphasizing potential terms in the corpus. Moreover, the soft clustering approach causes long-tail noise by putting every word into every cluster, which affects the natural thematic representation of documents and their proper classification. It is more challenging to capture semantic insights when dealing with short-length documents where word co-occurrence information is limited.

Keywords: classification ; text mining ; natural language processing

1. Introduction

Language modeling with binary one-hot word encoding is higher dimensional and sparse with no semantic information. As a result, the word analogy is missing; e.g., the distance between word vectors represents only the difference in alphabetic ordering. However, point vector representation of words in the embedding space like word2vec [1] and GloVe [2][3] contain semantic information. Representing a dense low-dimensional fixed-length document vector is much more expensive and complicated. Moreover, it is challenging to infer unseen documents during the test process. The simplest method to get document embedding is the weighted averaging of word embeddings in the document [4]. Document Vector through Corruption (Doc2VecC) [5] is a significant study that shows how a simple weighted averaging technique combined with a simple noise-eliminating procedure can be effective. However, Doc2VecC did not cover the underlying themes of the document. Sparse Composite Document Vector (SCDV) [6] addresses this limitation by using word embeddings and the Gaussian mixture clustering to generate the document vector, which also overcomes the shortcomings of the hard-clustering approach [7]. SCDV shows significant improvement in the downstream natural language processing (NLP) tasks, including document classification. However, SCDV inherits noisy tails from the Gaussian mixture clustering that is not appropriate for the document containing multiple sentences [8]. Another vital issue ignored by most document representation methods is ignoring potential terms in the corpus, which is essential for understanding deep semantic insight of the documents.

It is hard to encode the richness of semantic insights for short-length documents where word co-occurrence information is limited [9]. Therefore, many works suggest importing semantic information from external sources [10][11][12]. However, accessing information from external sources (e.g., Wikipedia) may cause irrelevant noise corresponding to the current short text corpus.

Sparse Composite Document Vector with Multi-Sense Embeddings (SCDV-MS) [13] forces discarding the outliers from the clustering output to eliminate long-tail noises in the SCDV, which applies a hard threshold that may hinder the thematic representation of documents. Moreover, representing proper expressive documents depends upon modeling the underlying semantic topics in the correct form [14], which requires capturing deep semantics insights buried in words, expressions, and string patterns [15]. Hence, for the noisy long texts, researchers proposed *Weighted Sparse Document Vector* (WSDV) that embodies important words emphasizing capability using *Pomegranate General Mixture* model [16], and a soft threshold-based noise reduction technique.

It is challenging to capture semantics insights in document modeling with sparse short texts. The probability distribution of words captures better semantics than the point embedding approach (e.g., word2vec) [17] as it generalizes deterministic point embeddings of the terms using the mean vector, and the covariance matrix holds uncertainty of the estimations. Hence, instead of depending on external knowledge sources, researchers proposed corpus statistics empowered *Weighted Compact Document Vector* (WCDV), which emphasizes potential terms while performing probability word distribution using the weighted energy function. In WCDV, researchers employ the *Multimodal word Distributions* [18] that learns distributions of words using the Expected Likelihood Kernel [19], which computes the inner product between distributions of words to get the affinity of word pairs. However, every word in a document does not hold the same

importance; some are used more frequently than others, indicating their importance in the corpus. It is required to emphasize the frequently used words, especially when word co-occurrence information is limited (e.g., microblogging, product review, etc.). Therefore, to preserve the word frequency importance, researchers proposed *Weight attained Expected Likelihood Kernel* which considers term frequency-based point weights while measuring the partial log energy between distributions in the *Multimodal word Distributions* [18].

2. Corpus Statistics Empowered Document Classification

Word embeddings models ignore side information (e.g., document labels) while learning embeddings from enormous document corpora. To improve word representation and text classification accuracy, Linear, Y. et al. [20] proposed to use document labels as the global context both in the local neural network model and the global matrix factorization framework. Obayes, H.K. et al. [21] combined GloVe and bidirectional long short-term memory (BiLSTM) recurrent neural network for better sentiment classification, which causes expensive computation and no guidance for documents containing multiple sentences. Yang, Z. et al. [22] proposed Hierarchical attention networks (HAN) for document classification, which maintain a hierarchical structure of word to sentence (building sentence from words) and sentence to document (aggregating sentences to a document representation). Zhang, Z. et al. [23] proved that the TFIDF algorithm with the combination of Naive Bayes has significance in the text classification task compared to many complex models.

Recently, transformers-based models [24][25] became more prevalent in downstream Natural Language Processing (NLP) tasks (e.g., document classification). Wang, B. andlinebreak Kuo, C-C.J. [26] proposed SBERTWK for sentence embedding, which trains on both word and sentence level objectives but no guidance for representing a document that contains multiple sentences. However, the transformer-based model requires enormous computational resources. Sanh, V. et al. [27] introduced a distilled version of BERT called DistilBERT, which is smaller, faster, cheaper, and lighter than other transformers-based models.

Mapping sentences to a fixed-length embedding vector using Universal Sentence Encoder (USE) based method [28] also got success in the downstream Natural Language Processing (NLP) task. The sentence analysis method made by combining Universal Language Model Fine-tuning (ULMFiT) with the Support Vector Machine (SVM) [29] is capable of performing document classification using a small amount of data but has higher computational complexity.

Yet, K.S. et al. [30] proposed document embedding along with their uncertainty called the Bayesian subspace multinomial model (Bayesian SMM) to capture better semantics. It is a generative log-linear model that learns to represent documents in the form of the Gaussian distributions and encodes uncertainty in the covariance matrix but holds only a single mode of words. Therefore, encoded uncertainty might diffuse spontaneously; the mean vector can be pulled in one direction and represents one particular meaning by leaving others not representing [31]. Different senses of a word lie in the linear superposition of standard word embeddings [32] and the Gaussian mixture model holds multiple modes to represent distinct meanings of words.

For the long texts classification, researchers proposed WCDV, which represents documents with uncertainty estimations in the distribution of words using Gaussian Mixtures distributions for short-length document classification. Researchers proposed WSDV using the Pomegranate General Mixture model for the long texts classification. Both WSDV and WCDV accommodate polysemous terms and train on the labeled documents corpus for better classification performance.

Noisy topics are outliers prone, thus less coherent and less expressive. Newman, D. [33] regularized the LDA-based topic model where only the higher frequency terms allow into the word dependencies sparse covariance matrix. This model executes two prime steps. Firstly, measuring the point weight of each word in the vocabulary, and secondly, putting a threshold point to eliminate lower weighted words from the covariance matrix. Mittal, M. et al. [34] introduced automated K-means clustering, where they applied a threshold point to decide whether or not to create a new cluster for the objects. This approach prohibits outlier tendency by accommodating lower probability objects into a new cluster. Gupta, V. et al. [13] introduced SCDV-MS, which removes noise by applying a hard threshold on the fuzzy word cluster assignments, which proved better classification performance and lower space and time complexity than SCDV [6].

In contrast, the proposed WSDV contains more natural noise removal techniques using a soft threshold and more efficient sparse vectorial representation for the long text (e.g., removing first principle components).

To capture better corpus semantics, Sia, S. et al. [35] introduced weighted data clustering on pre-trained word embeddings, where they also proved the effectiveness of re-ranking the top words in a cluster for better representative topics. Similarly, Gebru, I.D. et al. [36] proposed a Gaussian mixture-based weighted data clustering method called WD-GMM that demonstrates how the point weight of datum affects the covariance matrix and leads to better clustering.

Inspired by them, researchers proposed WSDV, which extends the clustering process on the weighted data for the multi-class document classification performance.

Short texts are sparse due to limited word co-occurrence, which requires special treatment to capture hidden semantic information [37][38]. Pretrained word embedding over large external corpora is a common remedy for dealing short length documents. Zuo, Y. et al. [39] proposed a word embedding-enhanced Pseudo-document-based Topic Model (WE-PTM) to leverage pre-trained word embeddings that is essential for alleviating data sparsity. Instead of incorporating external knowledge sources, Zhang, P. and He, Z. [40] proposed an ensemble approach by exploiting both word embeddings and latent topics in sentence-level sentiment analysis for sentence polarity detection.

Therefore, for semantically enriched short-length document representation, instead of importing information from external knowledge sources, researchers employ *Multimodal word Distributions* [18] to capture uncertainty in the distribution of word embeddings for the vectorial representation of documents.

The contextual analysis-based model emphasizes potential terms that capture better semantics insights and boost classification performance [41]. Xu, J. et al. [42] proposed a convolutional neural network-based model, which incorporates context-relevant concepts into text representation for uplifting short text classification performance, but it requires expensive computational capacity.

In WCDV, researchers use the weighted energy function to emphasize potential terms in the short texts corpus.

Weighted Kernel Density Estimation (WKDE) [43][44] based on point weights has proved effective. For the semantic similarity measuring task, constant weighting assumption-based semantic similarity [45] measure between two concepts/words holds better performance for the semantic representation of the concept/words but holds the same weighting relevance. Later, it found that the weight propagation mechanism [46][47] for augmenting input with semantic information achieves desired performance and removes the same weighting curse for concepts/words. Recently, Liu J. et el. [48] introduced a weighted kernel mechanism for the weighted k-means multi-view clustering, where they redefined the objective by assigning weights to the cluster level instead of global weighting for each view and outperforming the existing objective.

References

1. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 2013, 26.
2. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
3. Wang, H. Extremal GloVe: Theoretically Accurate Distributed Word Embedding by Tail Inference. In Proceedings of the 7th International Conference on Communication and Information Processing (ICCIP), Beijing, China, 22–24 October 2021; pp. 1–3.
4. Kusner, M.J.; Sun, Y.; Kolkin, N.I.; Weinberger, K.Q. From word embeddings to document distances. In Proceedings of the ICML, Lille, France, 7–9 July 2015.
5. Minmin, C. Efficient vector Representation for Documents through Corruption. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
6. Gupta, V.; Mekala, D.; Paranjape, B.; Karnick, H. Scdv: Sparse composite document vectors using soft clustering over distributional representations. In Proceedings of the EMNLP, Austin, TX, USA, 1–5 November 2016.
7. Gupta, V.; Karnick, H.; Bansal, A.; Jhala, P. Product classification in e-commerce using distributional semantics. In Proceedings of the 26th International Conference on Computational Linguistics, Osaka, Japan, 11–17 December 2016; pp. 536–546.
8. Gupta, V.; Saw, A.; Nokhiz, P.; Netrapalli, P.; Rai, P.; Talukdar, P. P-SIF: Document Embeddings Using Partition Averaging. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown, New York, NY, USA, 7–12 February 2020; pp. 7863–7870.
9. Lovera, F.A.; Cardinale, Y.C.; Homsi, M.N. Sentiment Analysis in Twitter Based on Knowledge Graph and Deep Learning Classification. Electronics 2021, 10, 2739.

10. Weng, J.; Lim, E.; Jiang, J.; He, Q. Twitterrank: Finding topic-sensitive influential twitterers. In Proceedings of the third ACM International Conference on Web Search and Data Mining, New York, NY, USA, 3–6 February 2010; pp. 261–270.
11. Phan, X.; Nguyen, L.; Horiguchi, S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In Proceedings of the 17th International Conference on World Wide Web, Beijing, China, 21–25 April 2008; pp. 91–100.
12. Yi, F.; Bo, J.; Wu, J. Topic modeling for short texts via word embedding and document correlation. IEEE Access 2020, 8, 30692–30705.
13. Gupta, V.; Saw, A.; Nokhiz, P.; Gupta, H.; Talukdar, P. Improving document classification with multi-sense embeddings. In Proceedings of the 24th European Conference on Artificial Intelligence—ECAI, Santiago de Compostela, Spain, 29 August–8 September 2020.
14. Liu, Y.; Liu, Z.; Chua, T.; Sun, M. Topic modeling for sequential documents based on hybrid inter-document topic dependency. J. Intell. Inf. Syst. 2021, 56, 435–458.
15. Šuman, S.; Čandrlić, S.; Jakupović, A. A Corpus-Based Sentence Classifier for Entity–Relationship Modelling. Electronics 2022, 11, 889.
16. Schreiber, J. Pomegranate: Fast and Flexible Probabilistic Modeling in Python. J. Mach. Learn. Res. 2017, 18, 5992–5997.
17. Navigli, R.; Martelli, F. An overview of word and sense similarity. Nat. Lang. Eng. 2020, 25, 693–714.
18. Athiwaratkun, B.; Wilson, A.G. Multimodal word distribution. In Proceedings of the 55th Annual Meeting of the ACL, Vancouver, BC, Canada, 30 July–4 August 2017; Volume 1, pp. 1645–1656.
19. Jebara, T.; Kondor, R. Bhattacharyya and expected likelihood kernels. Learn. Theory Kernel Mach. 2003, 57–71.
20. Yang, L.; Chen, X.; Liu, Z.; Sun, M. Improving Word Representations with Document Labels. IEEE/ACM Trans. Audio Speech Lang. Process. 2017, 25, 863–870.
21. Obayes, H.K.; Al-Turaihi, F.S.; Alhussayni, K.H. Sentiment classification of user's reviews on drugs based on global vectors for word representation and bidirectional long short-term memory recurrent neural network. Indonesian J. Electr. Eng. Comput. Sci. 2021, 23, 345–353.
22. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical attention networks for document classification. In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
23. Zhang, Z.; Wu, Z.; Shi, Z. An improved algorithm of TFIDF combined with Naive Bayes. In Proceedings of the 7th International Conference on Multimedia and Image Processing, Suzhou, China, 20–22 July 2022; pp. 167–171.
24. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M. Transformers: State-of-the-art natural language processing. In Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.
25. Bilal, M.; Almazroi, A.A. Effectiveness of Fine-Tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews. Electron. Commer. Res. 2022, 38–45.
26. Wang, B.; Kuo, C.-C.J. SBERT-WK: A Sentence Embedding Method by Dissecting BERT-Based Word Models. IEEE/ACM Trans. Audio Speech Lang. Process. 2020, 28, 2146–2157.
27. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv 2019, arXiv:1910.01108.
28. Pradhan, R.; Sharma, D.K. An ensemble deep learning classifier for sentiment analysis on code-mix Hindi—English data. Soft Comput. 2022, 1–18.
29. AlBadani, B.; Shi, R.; Dong, J. A novel machine learning approach for sentiment analysis on Twitter incorporating the universal language model fine-tuning and SVM. Appl. Syst. Innov. 2022, 5, 13.
30. Kesiraju, S.; Plchot, O.; Burget, L.; Suryakanth, V.G. Learning Document Embeddings Along With Their Uncertainties. IEEE/ACM Trans. Audio Speech Lang. Process. 2020, 28, 2319–2332.
31. Chen, X.; Qiu, X.; Jiang, J.; Huang, X. Gaussian Mixture Embeddings for Multiple Word Prototypes. arXiv 2015, arXiv:1511.06246.
32. Arora, S.; Li, Y.; Liang, Y.; Ma, T.; Risteski, A. Linear algebraic structure of word senses, with applications to polysemy. Trans. Assoc. Comput. Linguist. 2018, 6, 483–495.

33. Newman, D.; Bonilla, E.V.; Buntine, W. Improving Topic Coherence with Regularized Topic Models. Adv. Neural Inf. Process. Syst. 2011, 24, 496–504.
34. Mittal, M.; Sharma, R.K.; Singh, V.P. Validation of k-means and Threshold based Clustering Method. Int. J. Adv. Technol. 2014, 5, 153–160.
35. Sia, S.; Dalmia, A.; Mielke, S.J. Tired of topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! arXiv 2020, arXiv:2004.14914.
36. Gebru, I.D.; Alameda-Pineda, X.; Forbes, F.; Horaud, R. EM Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 2402–2415.
37. Chen, L.M.; Xiu, B.; Ding, Z.Y. Multiple weak supervision for short text classification. Appl. Intell. 2022, 52, 9101–9116.
38. Murakami, R.; Chakraborty, B. Neural Topic Models for Short Text Using Pretrained Word Embeddings and Its Application to Real Data. In Proceedings of the 4th International Conference on Knowledge Innovation and Invention (ICKII), Taichung, Taiwan, 23–25 July 2021; pp. 146–150.
39. Zuo, Y.; Li, C.; Lin, H.; Wu, J. Topic modeling of short texts: A pseudo-document view with word embedding enhancement. IEEE Trans. Knowl. Data Eng. 2021.
40. Zhang, P.; He, Z. Using data-driven feature enrichment of text representation and ensemble technique for sentence-level polarity classification. J. Inf. Sci. 2010, 41, 531–549.
41. Sulaimani, S.; Starkey, A. Multiple weak supervision for short text classification. IEEE Access 2022, 9, 149619–149629.
42. Xu, J.; Cai, Y.; Wu, X.; Lei, X.; Huang, Q.; Leung, H.; Li, Q. Incorporating context-relevant concepts into convolutional neural networks for short text classification. Neurocomputing 2020, 386, 42–53.
43. Fieberg, J. Utilization distribution estimation using weighted kernel density estimators. J. Wildl. Manag. 2007, 71, 1669–1675.
44. Zhou, H.; Cheng, Q.; Yang, H.; Xu, H. Weighted Kernel Density Estimation of the Prepulse Inhibition Test. In Proceedings of the 6th World Congress on Services, Miami, FL, USA, 5–10 July 2010; pp. 291–297.
45. Saif, A.; Zainodin, U.Z.; Omar, N.; Ghareb, A.S. Weighting-based semantic similarity measure based on topological parameters in semantic taxonomy. Nat. Lang. Eng. 2018, 24, 861–886.
46. Pittaras, N.; Giannakopoulos, G.; Papadakis, G.; Karkaletsis, V. Text classification with semantically enriched word embeddings. Nat. Lang. Eng. 2020, 27, 391–425.
47. Yue, T.; Li, Y.; Hu, Z. DWSA: An Intelligent Document Structural Analysis Model for Information Extraction and Data Mining. Electronics 2021, 10, 2443.
48. Liu, J.; Cao, F.; Gao, X.; Yu, L.; Liang, J. A Cluster-Weighted Kernel K-Means Method for Multi-View Clustering. In Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown, New York, NY, USA, 7–12 February 2020; pp. 4860–4867.