Unsupervised Contextualized Document Representation

Ankur Gupta
Indian Institute of Technology Kanpur

Vivek Gupta
School of Computing, University of Utah

Second Workshop on Simple and Efficient Natural Language Processing (SustaiNLP 2021), EMNLP 2021

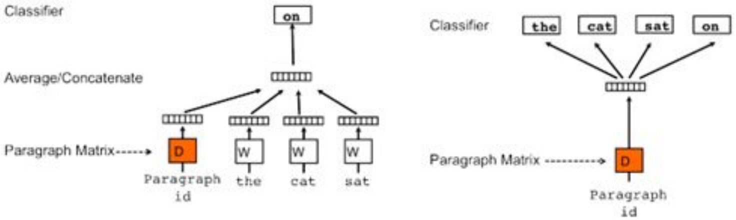
Motivation

- Natural language requires good semantic representations of **textual documents**
    - Text Categorization
    - Information Retrieval
    - Text Similarity

- Good semantic representation of words exists, i.e., **Word2vec (SGNS, CBOW)** created by Mikolov et al., **Glove** (Socher et al.) and many more.

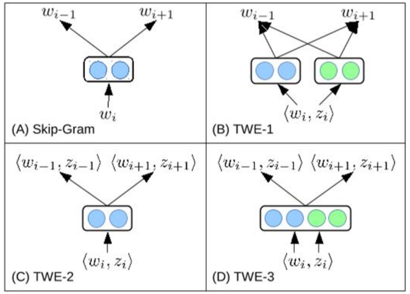
- **What About Documents?**
    - **Multiple Approaches** based on **local context, topic modelling, context sensitive learning**
    - **Semantic Composition** in natural language is the task of modelling the meaning of a larger piece of text *(document)* by composing the meaning of its constituents/parts *(words).*
        - *Our work focus on using simple semantic composition*

Efforts for Document Representation

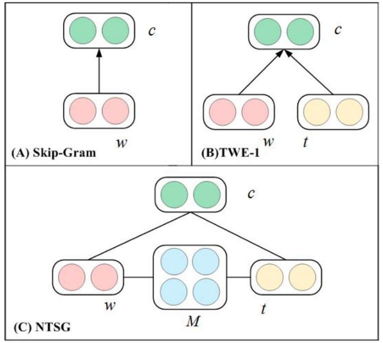


Doc2Vec (Le & Mikolov, 2014)
Local + Global context

Deep Learning
LSTM, RNN, Bi-LSTM,
RTNN, LSTM Attention
Contextual Embedding
ELMo, BERT



TWE (Liu et al., 2015a)
Topic Modelling



NTSG (Liu et al., 2015b)
Topic Modelling + Context Sensitive Learning

~~Larger Document Multiple topic~~

Graded Weighted M
2015, Arora
Weighted Average o position

Sentence Embedding

Our Approach:
SCDV+BERT(cxtd)

↓

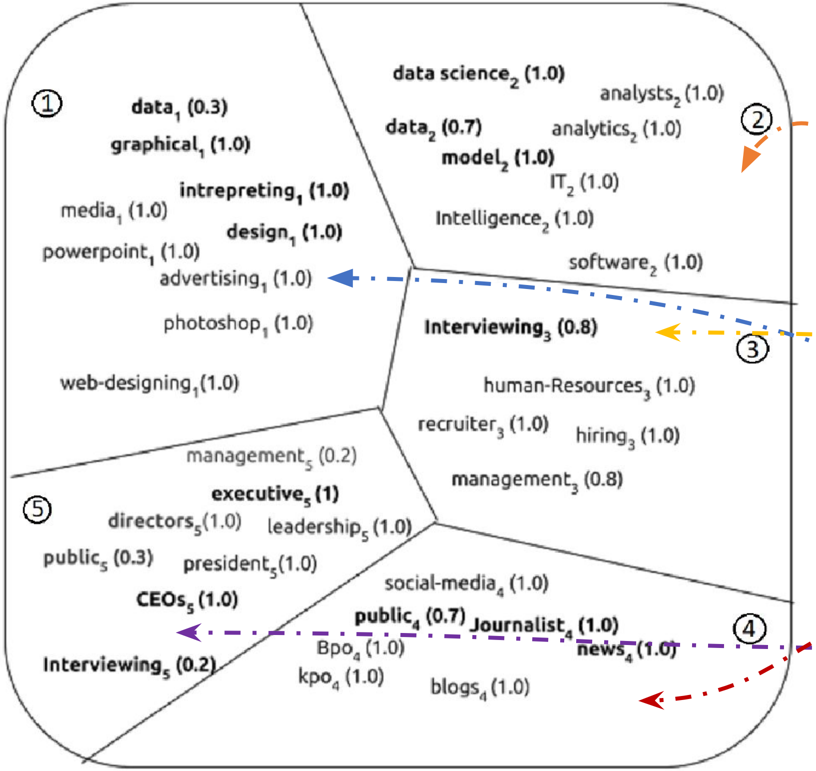
BERT
(Document Contextualization)

\+

SCDV
(Sparse Contextualized Document Vector)

SCDV : Averaging vs Partition Averaging

"Data journalists deliver the news of data science to general public, they often take part in interpreting the data models, creating graphical designs and interviewing the director and CEOs."



SIMPLE AVERAGING

$$\vec{v}_{\text{data}_2} + \vec{v}_{\text{journalist}_4} + \vec{v}_{\text{news}_4} + \vec{v}_{\text{datascience}_1} + \vec{v}_{\text{public}_4} + \vec{v}_{\text{interpreting}_1} + \vec{v}_{\text{models}_2} + \vec{v}_{\text{graphical}_1} + \vec{v}_{\text{design}_1} + \vec{v}_{\text{director}_5} + \vec{v}_{\text{CEO}_5} + \vec{v}_{\text{interviewing}_2}$$

Here, $\oplus$ represent concatenation, and + represent addition

WEIGHTED PARTITION AVERAGING

$$(\vec{v}_{\text{interpreting}_1} + \vec{v}_{\text{graphical}_1} + \vec{v}_{\text{design}_1} + 0.3 * \vec{v}_{\text{data}_1}) \oplus (0.7 * \vec{v}_{\text{data}_2} + \vec{v}_{\text{datascience}_2} + \vec{v}_{\text{models}_2}) \oplus (\vec{v}_{\text{journalist}_4} + \vec{v}_{\text{news}_4} + 0.7 * \vec{v}_{\text{public}_4}) \oplus (\vec{v}_{\text{director}_5} + 0.3 * \vec{v}_{\text{public}_5} + \vec{v}_{\text{CEO}_5} + 0.2 * \vec{v}_{\text{interviewing}_5}) \oplus 0.8 * \vec{v}_{\text{interviewing}_3}$$

+ within a partition and $\oplus$ across partitions

Document Contextualization (Using BERT)

→ Classify the following sentences according to used context:

The **stocks** of apple have increased. ⟶ **Company** ✅

For **health life**, eat an **apple** everyday. ⟶ **Company** ✘

Raw Docs

Messi scored a penalty!
Judge passed the order of …
The court issued a penalty ….

Contextualized Docs

Messi scored a **penalty$1**!
Judge passed the order of …
The **court$1** issued a **penalty$0** ….

Document Contextualization (Using BERT)

- Use **K-Means** algorithm (Jain and Dubes, 1988) to cluster all the **BERT contextualized** representations of **all occurance of the word**.

- Cluster centers are **symbolic representations** of several **meaning of a word** in different contexts it can occur in the corpus across documents**.**

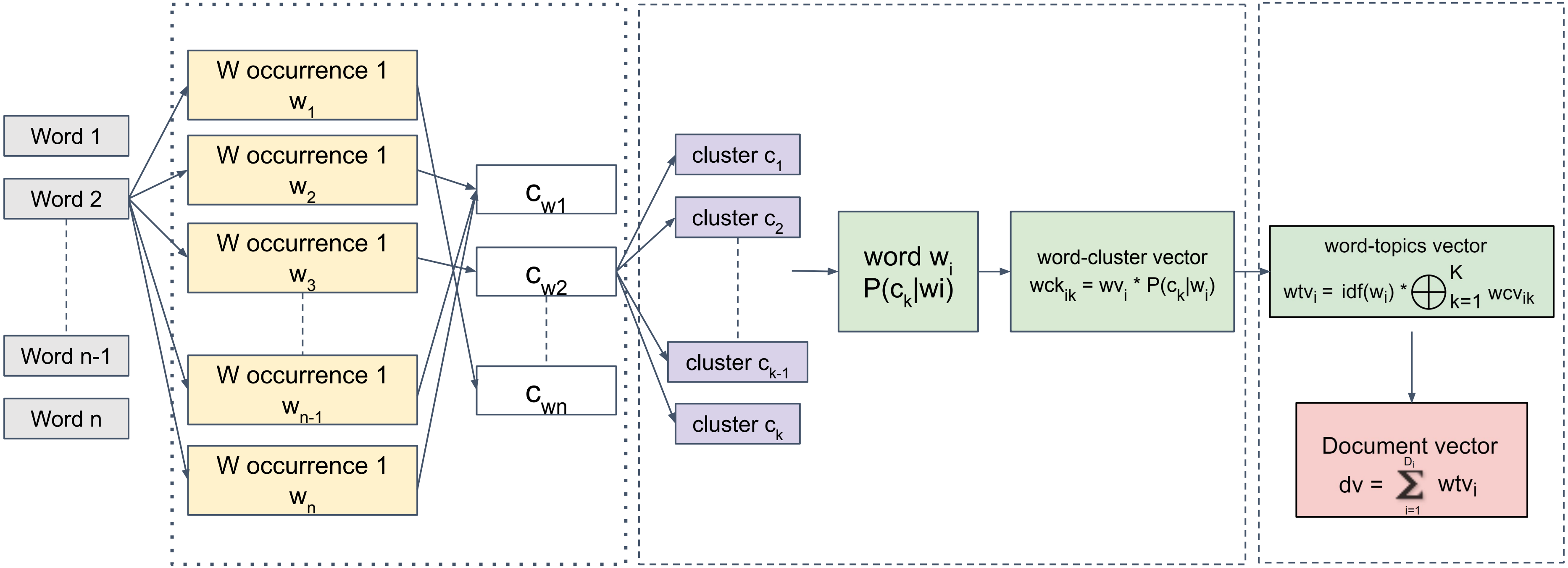
- Why K-Means?
    - It is **efficient,** needed as **clustering** for **all vocabulary words**.
    - **Cosine similarity = Euclidean distance** for unit norm vectors.

- Cluster similarity **threshold is a hyperparameter** of our algorithm.

Corpus Contextualization (WSD)

Word Cluster Vector Formation (SCDV)

Final Document Representation



This process occurs for each unique word in the corpus

This process occurs for each disambiguated word in the corpus

Summation wtv for word in doc.

Multi-Class Classification – 20NewsGroup (40-80 words)

Model	Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)
SCDV+BERT (cxtd) + Anisotropy	**86.9**	**86.4**	**86.1**	**86.3**
SCDV + word2vec	84.6	84.6	84.5	84.6
BERT (pr)	84.9	84.9	85.0	85.0
BoWV	81.6	81.1	81.1	80.9
weight -Avg (SIF)	81.9	81.7	81.9	81.7
NTSG-1	82.6	82.5	81.9	81.2
TWE-1	81.5	81.2	80.6	80.6
Doc2Vec	75.4	74.9	74.3	74.3

Multi-Class Classification – 20NewsGroup (40-80 words)

Model	Accuracy (↑)	Precision (↑)	Recall (↑)	F1-Score (↑)
SCDV+BERT (cxtd) + Anisotropy	**86.9**	**86.4**	**86.1**	**86.3**
SCDV + word2vec	84.6	84.6	84.5	84.6
BERT (pr)	84.9	84.9	85.0	85.0

Performance of **SCDV + BERT(cxtd) + Anisotropy** better the baselines

- SCDV + word2vec
- BERT (pre-trained)

Show that **both BERT base contextualization** and **SCDV** are **important**

Multi-Class Classification – Comparison across Datasets

Dataset	BERT(pr)	SCDV + Word2Vec
Amazon	91.04	**93.9**
BBCSport	99.12	98.81
Twitter	66.63	**74.2**
Classic	95.63	**96.9**
Recipe-L	68.44	**78.5**
20NG	64.81	**84.9**

SCDV with word2vec embedding

performs better than

BERT(pr) pre-trained direct averaging.

→ except one dataset “*bbcsport*”

Multi-Class Classification – Comparison across Datasets

Dataset	BERT(pr)	SCDV + Word2Vec	SCDV + BERT(weight-avg)
Amazon	91.04	93.9	**94.62**
BBCSport	99.12	98.81	97.29
Twitter	66.63	**74.2**	72.98
Classic	95.63	**96.9**	96.54
Recipe-L	68.44	**78.5**	78.13
20NG	64.81	**84.9**	**84.9**

SCDV + BERT (weight-avg) performs almost similar to SCDV + Word2Vec

⟶

Simply Replacing Word2Vec with BERT gives no advantage

except on *"Amazon" datset*

Multi-Class Classification – Comparison across Datasets

Dataset	BERT(pr)	SCDV + Word2Vec	SCDV + BERT(weight-avg)	SCDV + BERT(ctxd)+ Anisotropy
Amazon	91.04	93.9	94.62	**95.88**
BBCSport	99.12	98.81	97.29	**99.60**
Twitter	66.63	74.2	72.98	**77.03**
Classic	95.63	96.9	96.54	**99.01**
Recipe-L	68.44	78.5	78.13	**80.74**
20NG	64.81	84.9	84.9	**86.94**

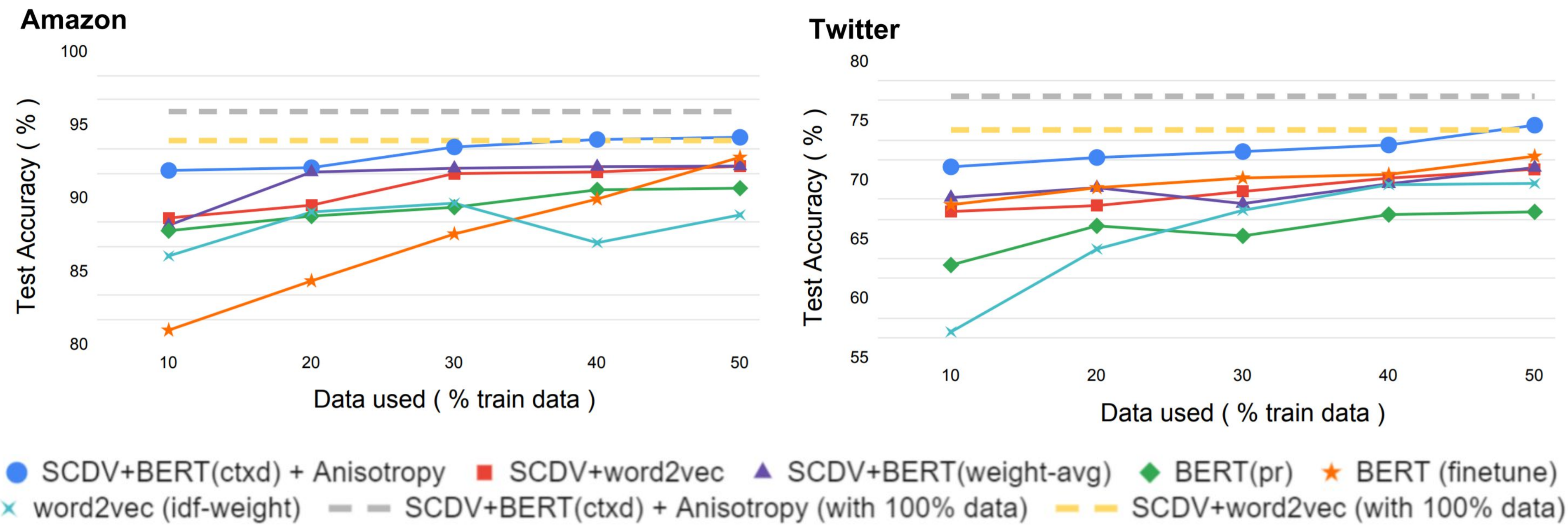
SCDV + BERT (ctxd) + Anisotropy outperforms SCDV + Word2Vec ⟶
Credits to BERT Contextualization (WSD) & Anisotropy adjustment.

Multi-Class Classification – Comparison across Datasets

Dataset	BERT(pr)	SCDV + Word2Vec	SCDV + BERT(weight-avg)	SCDV + BERT(ctxd)+ Anisotropy	BERT (finetune)
Amazon	91.04	93.9	94.62	**95.88**	94.6
BBCSport	99.12	98.81	97.29	99.60	**99.67**
Twitter	66.63	74.2	72.98	**77.03**	73.13
Classic	95.63	96.9	96.54	**99.01**	98.67
Recipe-L	68.44	78.5	78.13	80.74	**81.13**
20NG	64.81	84.9	84.9	**86.94**	86.91

SCDV + BERT (ctxd) + Anisotropy performs very similar / also sometime outperforms the BERT (finetune)

Embedding performance on Low Resource Setting

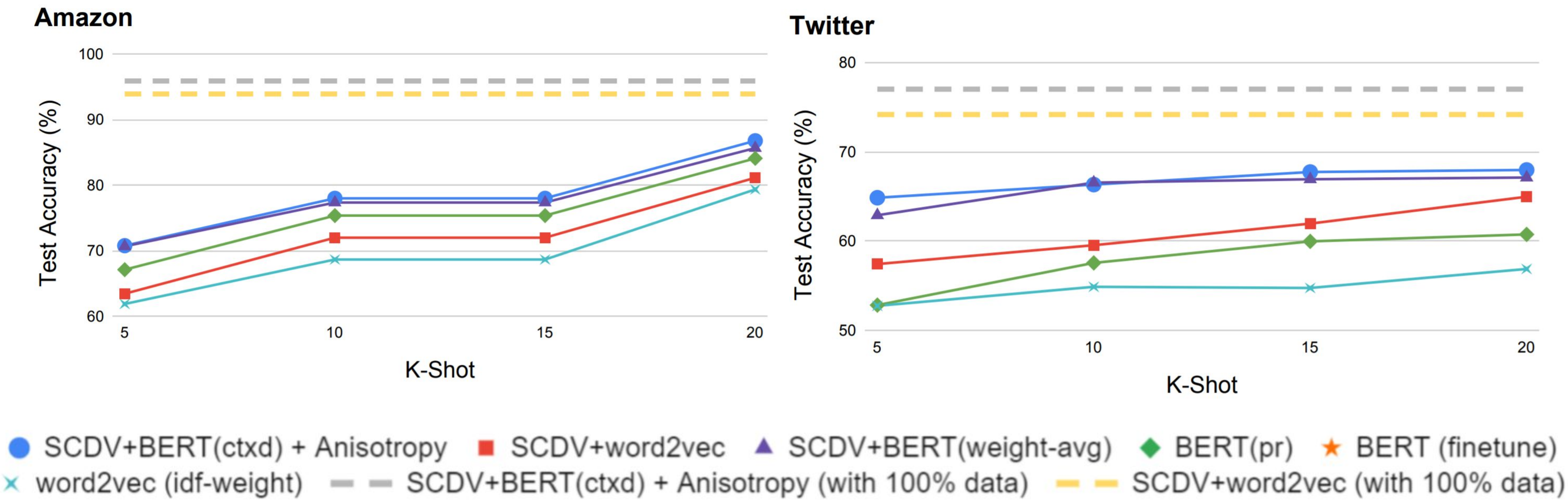


with limited training data

SCDV + BERT (ctxd) + Anisotropy **robust /stable** (with 40% > 100% SCDV+word2vec)

For BERT(fine-tune) **performance drop** drastically (even worse than SCDV+word2vec)

Embedding performance on Few-Shot Setting



SCDV + BERT (ctxd) + Anisotropy **outperform** BERT(fine-tune) and other models
All model improve performance with increasing K (#examples)

Semantic Textual Similarity (27 Datasets)

STS12	STS13	STS14	STS15	STS16
MSRpar	headline	deft forum	answers-forums	headlines
MSRvid	OnWN	deft news	answers-students	plagiarism
SMT-eur	FNWN	headline	belief	posteditng
OnWN	SMT	images	headline	answer-answer
SMT-news		OnWN	images	question-question
		tweet news		

Results (Pearson r X 100) on Semantic Textual Similarity

Model → Dataset ↓	PP -Proj	RNN	WME +PSL	Infer Sent	GRAN	Glove +WR	BERT (pr)	SCDV +w2v	SCDV + BERT (ctxd) + Anisotropy
STS12	60.0	58.4	62.8	61	62.5	56.2	53	59.5	**66.8**
STS13	56.8	56.7	56.3	56	63.4	56.6	**67**	61.8	64.1
STS14	71.3	70.9	68.0	68	**75.9**	68.5	62	73.5	**77.3**
STS15	74.8	75.6	64.2	71	**77.7**	71.7	73	76.3	**78.0**
STS16	-	64.9	-	**77**	-	72.4	67	72.5	74.6
Average	65.72	65.3	62.83	66.66	69.87	63.08	64.4	71.0	**72.22**

Concept Matching

The task is **to establish link** (<->) the **concept** with the **relevant projects.**

Concept Matching Dataset: **537 pairs (projects, concepts)**, **53 unique concepts** (NGSS) and **230 unique projects** from Science Buddies

Embedding	Accuracy	F1
TF-IDF	53.8	70.0
InferSent	54.0	70.1
BERT(pr)	54.8	70.6
SCDV + Word2Vec	53.7	70.0
SCDV + BERT(ctxd)	57.1	73.8
SCDV + BERT(ctxd) + Anisotropy	**58.9**	**74.6**

SCDV + BERT (ctxd) + Anisotropy

Outperform

SCDV + Word2Vec, BERT (pre-trained)

Takeaways

✓ Using **contextual representations** such as **BERT** for **word sense disambiguation** can lead to better document representations.

✓ SCDV's use of **partition-based averaging** rather than straight word vector averaging has a **significant influence** on document representation.

✓ **Anisotropic approach** for **isotropic reduction** are beneficial for getting better document representation, and hence the **corresponding downstream task**.

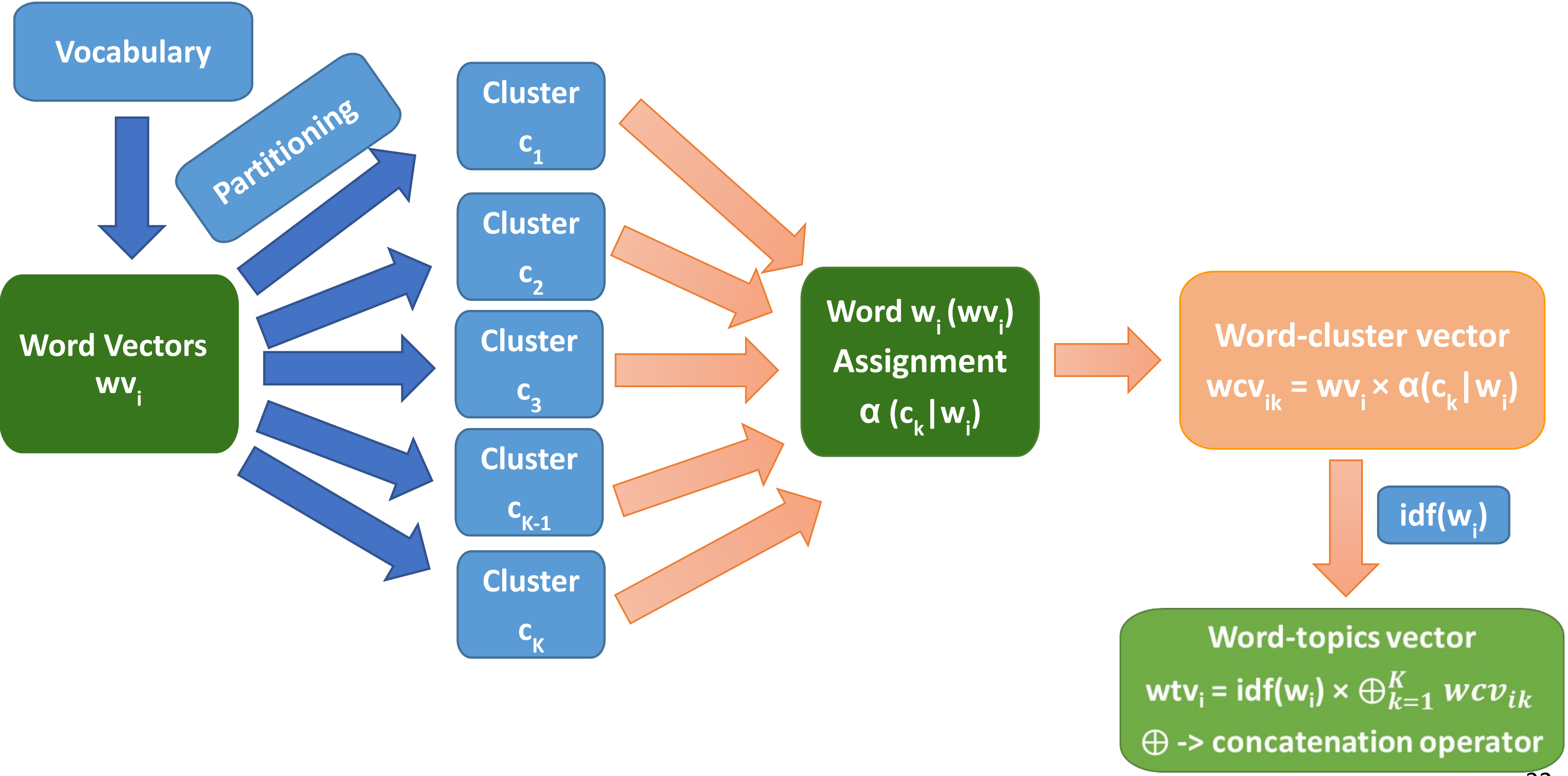
✓ **Fine tuning** of **contextual representation** such as **BERT not beneficial** for **low-resource setting with fewer labeled data**.

Paper : https://arxiv.org/pdf/2109.10509.pdf
Source : https://github.com/vgupta123/contextualize_scdv

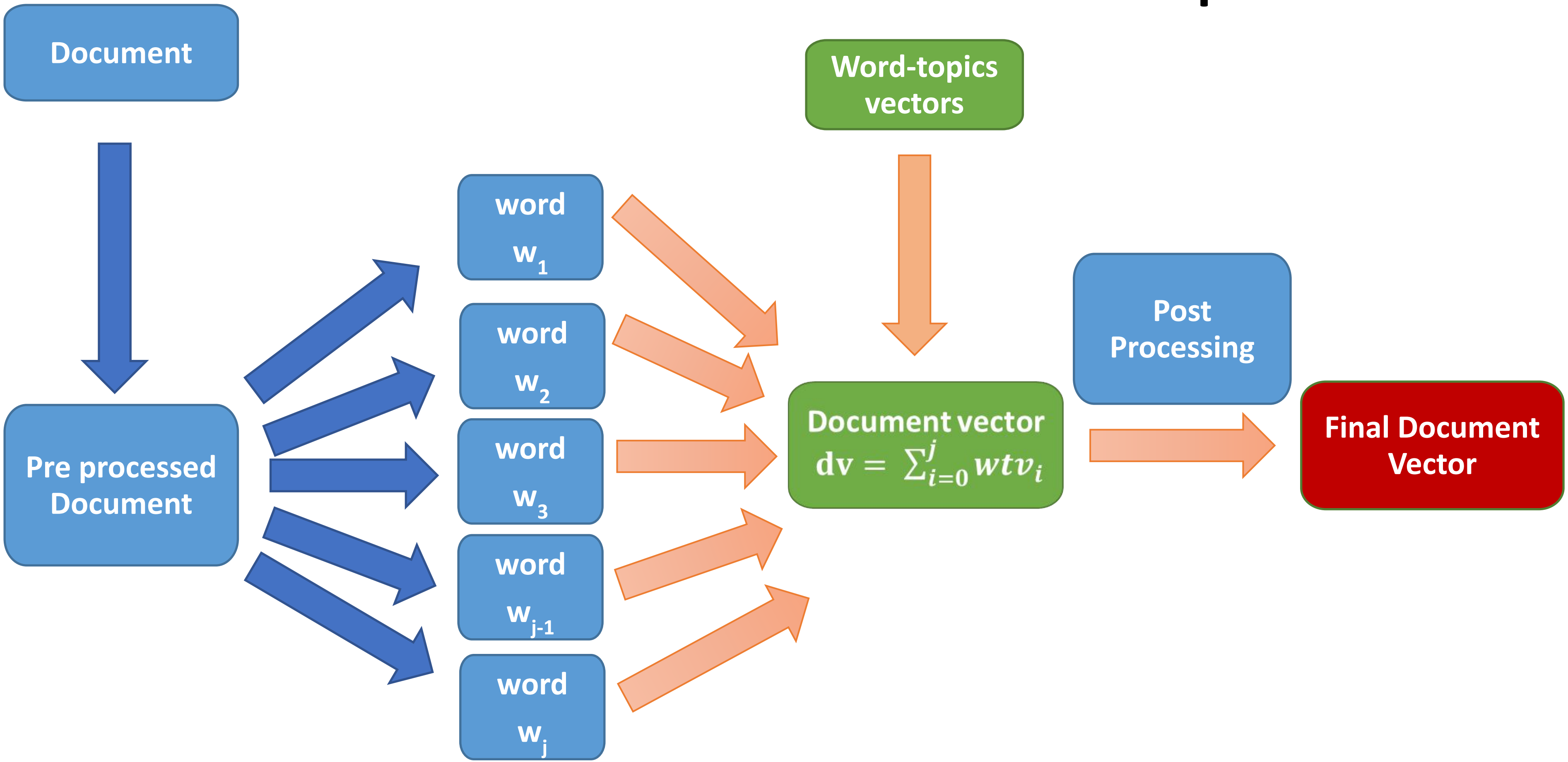
References

- **ConWea**: Dheeraj Mekala, Jingbo Shang, “*Contextualized Weak Supervision for Text Classification*”, In Proc ACL 2020
- **PSIF**: Vivek Gupta, Ankit Saw, Pegah Nokhiz, Praneeth Netrapalli, Piyush Rai, Partha Talukdar, “*P-SIF: Document Embeddings using Partition Averaging*”, In Proc AAAI 2020
- **BoWV** : Vivek Gupta and Harish Karnick et al, "*Product Classification in e-Commerce using Distributional Semantics*", In Proc COLING 2016
- **SCDV** : Dheeraj Mekala*,Vivek Gupta*, Bhargavi Paranjape and Harish Karnick, “*Sparse Composite Document Vectors using Soft Clustering over Distributional Semantics*”, In Proc EMNLP 2017
- **SCDV-MS** : Vivek Gupta et. al. “Word Polysemy Aware Document Vector Estimation”, In Proc ECAI 2020.
- **NTSG** : Pengfei Liu and Xipeng Qiu et al., *"Learning Context-Sensitive Word Embedding's with Neural Tensor Skip-Gram Model”*, In Proc IJCAI 2015
- **TWE** : Yang Liu and Zhiyuan Liu et al, *"Topical Word Embeddings”* In Proc AAAI, 2015
- **WMD** : Matt J. Kusner et al., *"From Word Embeddings To Document Distance*”, In ICML 2015
- **SIF** : Sanjeev Arora and Yingyu Liang “*A Simple but tough-to-beat baseline for sentence embedding's*”, In ICLR 2017
- **Polysemy** : Sanjeev Arora and Yuanzhi Li et al. “*Linear algebraic structure of word senses, with applications to polysemy”,* In TACL 2018
- **Doc2vec** *:* Quoc V Le and Tomas Mikolov. *“Distributed Representations of Sentences andDocuments*" In: ICML 2014

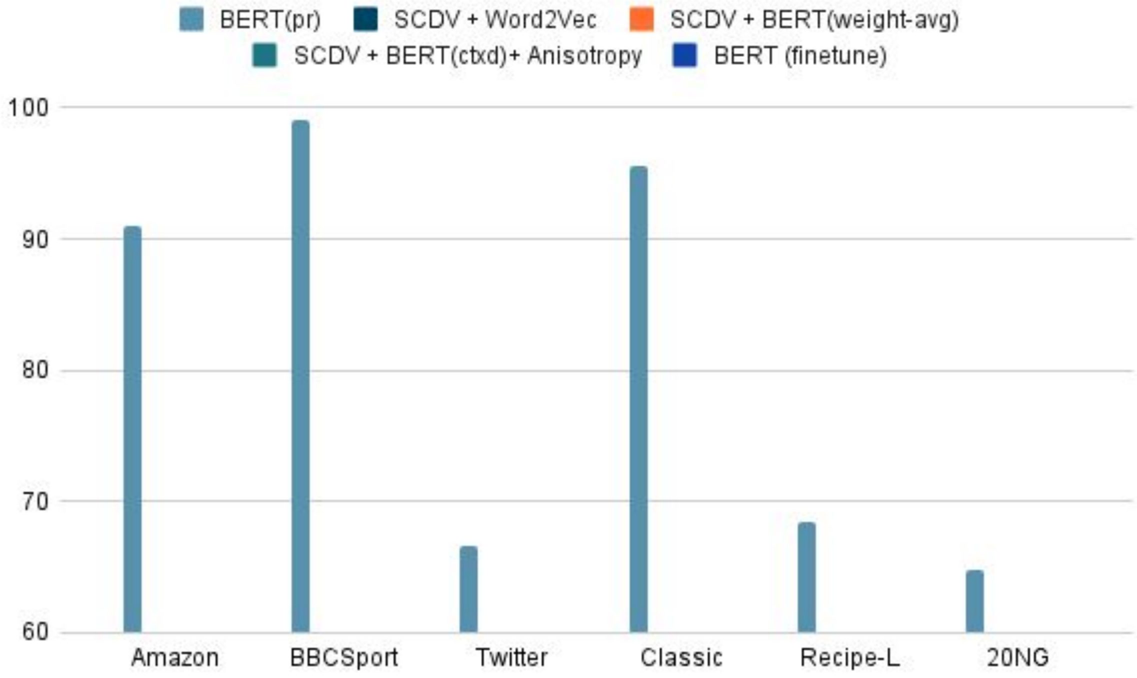
SCDV: Pre-computation of Word-topics Vector



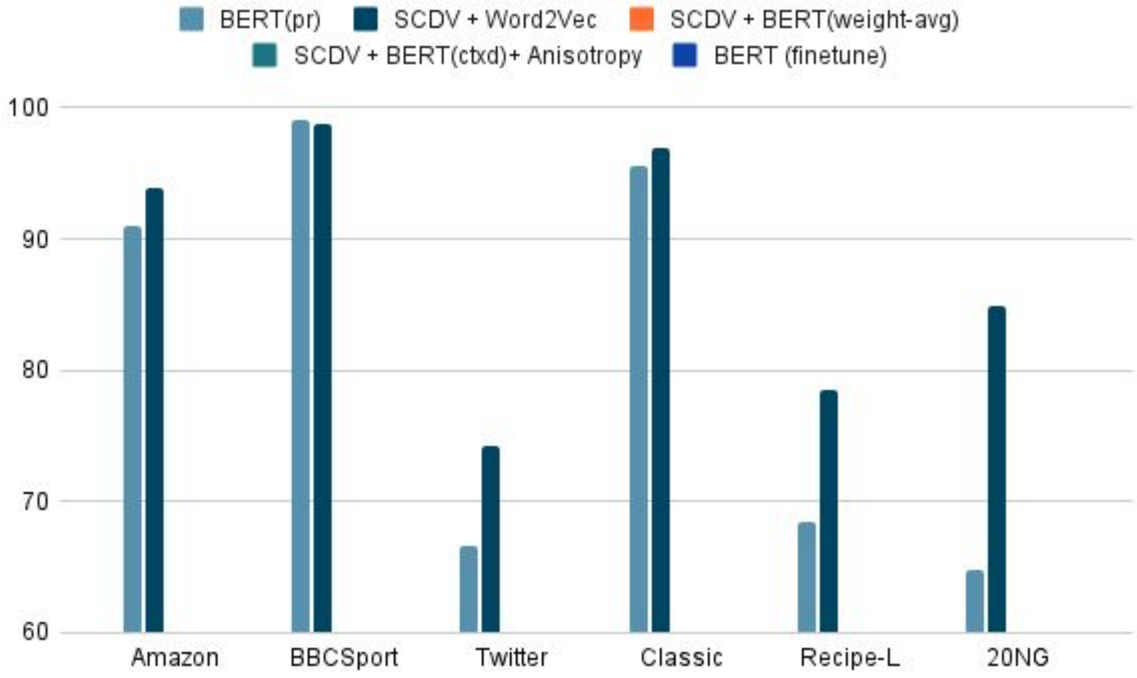
SCDV: Final Document Representation



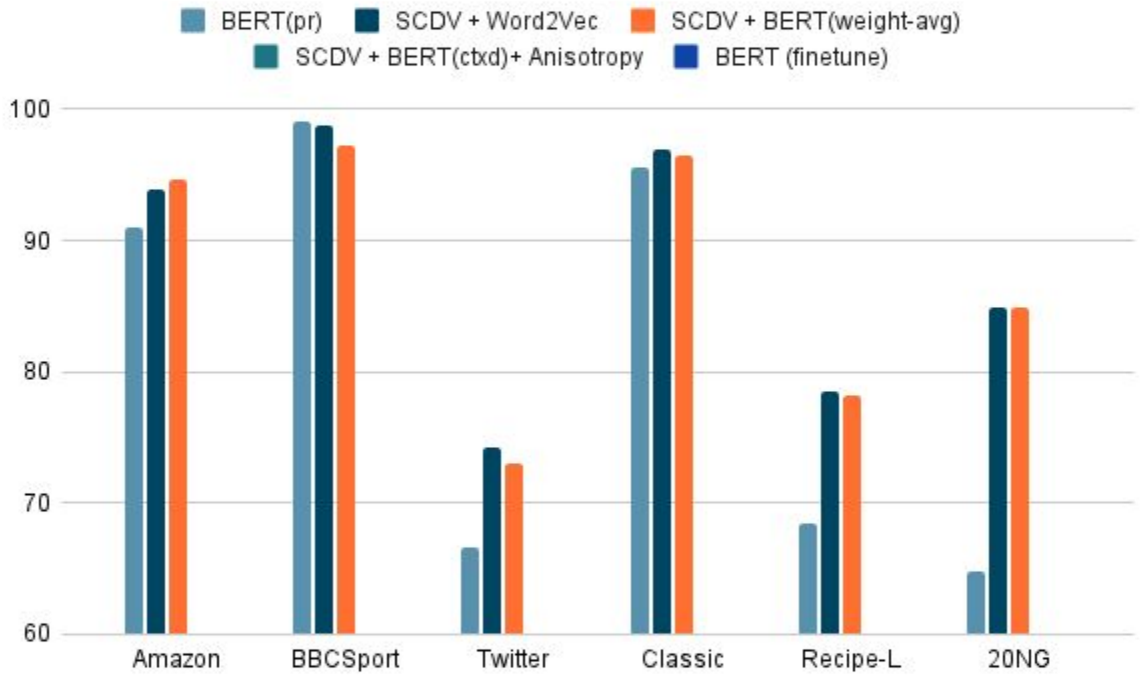
Embedding performance with complete training (full data setting)



Embedding performance with complete training (full data setting)



Embedding performance with complete training (full data setting)



Embedding performance with complete training (full data setting)

Embedding performance with complete training (full data setting)