Unsupervised Document Vector Representation using Partition Word-Vectors Averaging

Vivek Gupta[#’], Ankit Saw[*’], Partha Pratim Talukdar[^] and Praneeth Netrapalli[#]

[#] Microsoft Research Lab, India
[*]**Indian Institute of Technology, Kharagpur**
^Indian Institute of Science, Bangalore

24 January 2019

Data Science Seminar, Spring 2019

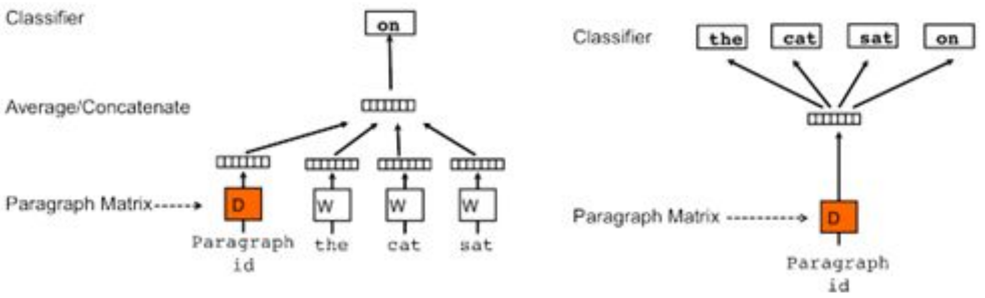
*** now Naukri.com (Infoedge India)**

‘ equal contribution

Motivation

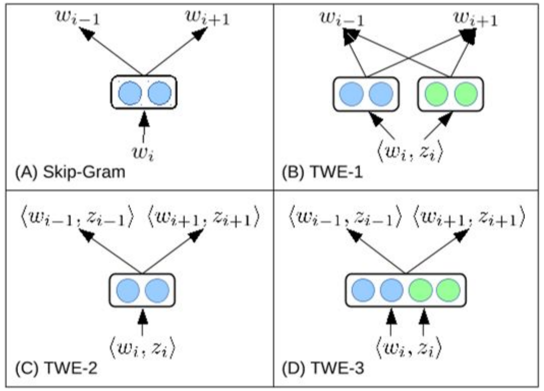
- Natural Language requires good semantic representations of **textual documents** for
  - Text Categorization
  - Information Retrieval
  - Sentiment Analysis
  - Text Similarity
- Good semantic representation of words exists i.e. **Word2vec (SGNS, CBOW)** created by Mikolov et al., **Glove** (Socher 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
(2014-2017)

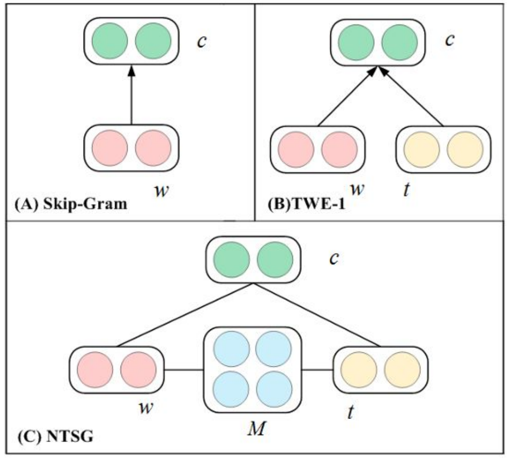


TWE (Liu et al., 2015a)
Topic Modelling

~~Larger Document Multiple topic~~

Graded Weighted Mode... ...s,
Arora et...
Weighted Avera... ...tion

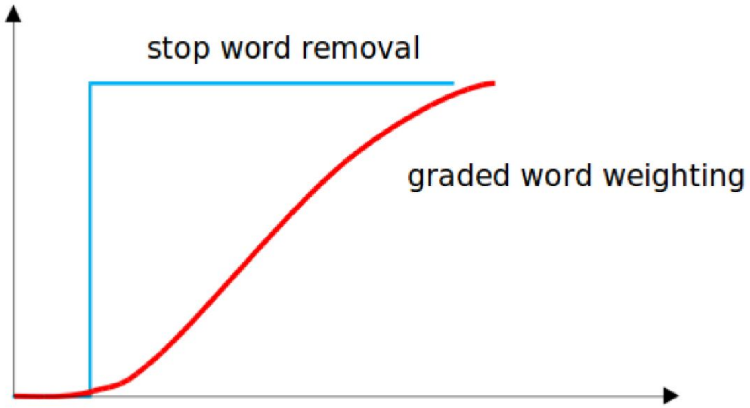
Sentence Embedding



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

Weighted Average of Word Vectors

- Varying weight capture the relative importance of the words
  - tf-idf weight
  - Smooth inverse frequency (SIF)

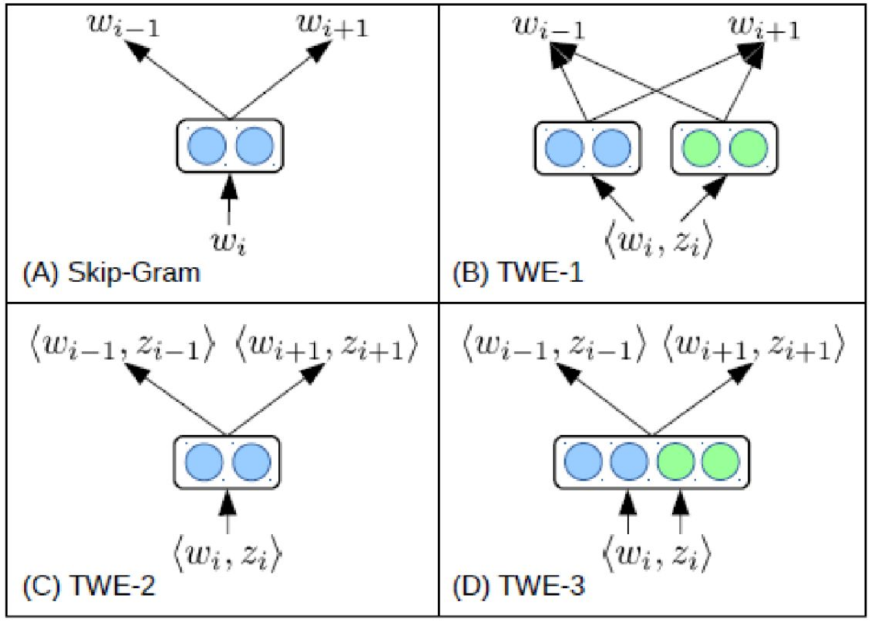


graded weighted by Mukerjee et. al.

- Arora et.al also applied PCA based post processing on vectors
  - Common Component Removal (WR)
- Work better than seq2seq model for representing a sentence

Background Semantic Composition

- Each word in corpus is assign to a topic using topic modeling algo (LDA)
- Four strategy was discussed to obtain the word and topic embeddings



- Document vectors … pic-word embedding weighted by tfidf

Topical Word Embedding (TWE)

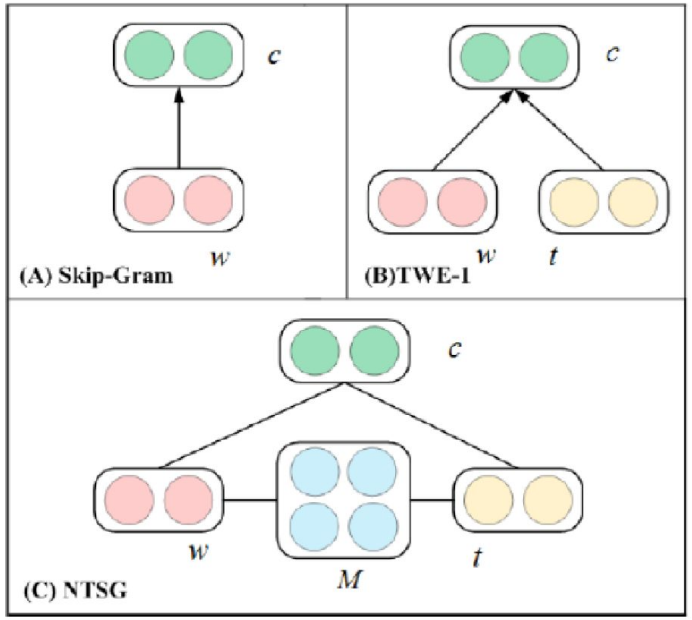
- TWE-1 learns word and topic embeddings by considering each topic as pseudo word (V + K vocabulary)
- TWE-2 directly consider each word-topic pair as a pseudo word (kV vocabulary, k is average active topic for each word)
- TWE-3 builds distinct embeddings for the topic and word separately and for each word-topic assignment, corresponding word embedding and topic embedding are concatenated after learning (weights are share)

Problems with TWE

- TWE-1 interaction between a word and the corresponding assign topic is not accounted.
- TWE-2, each word is differentiated into multiple topics which create sparsity and learning problems.
- TWE-3, the word embeddings are influenced by the corresponding topic embedding, making words in same topic less discriminative.
- TWE uses topic modelling algorithm like LDA to annotate words with topic, which make the feature formation slower
- Aggregating word-topic vectors to form document vectors average semantically different words.

Neural Tensor Skip Gram Model

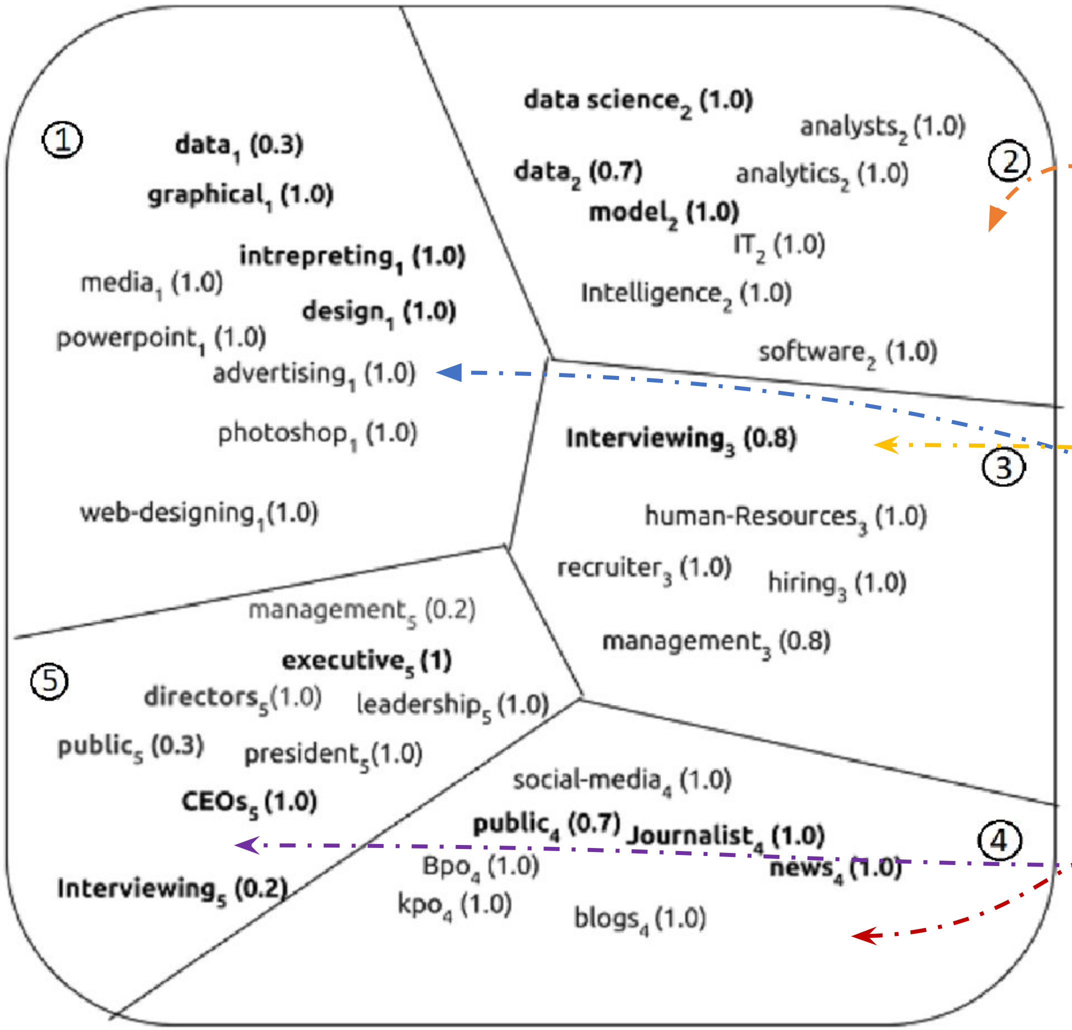
- TWE extension by learning a context sensitive word embeddings by using a tensor layer to model the interaction of words and topics.
- NTSG outperform majority embedding methods including TWE-1 on 20NewsGroup dataset



- Document vectors are ic-word embedding weighted by tfidf

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 CEO's.



SIMPLE AVERAGING

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

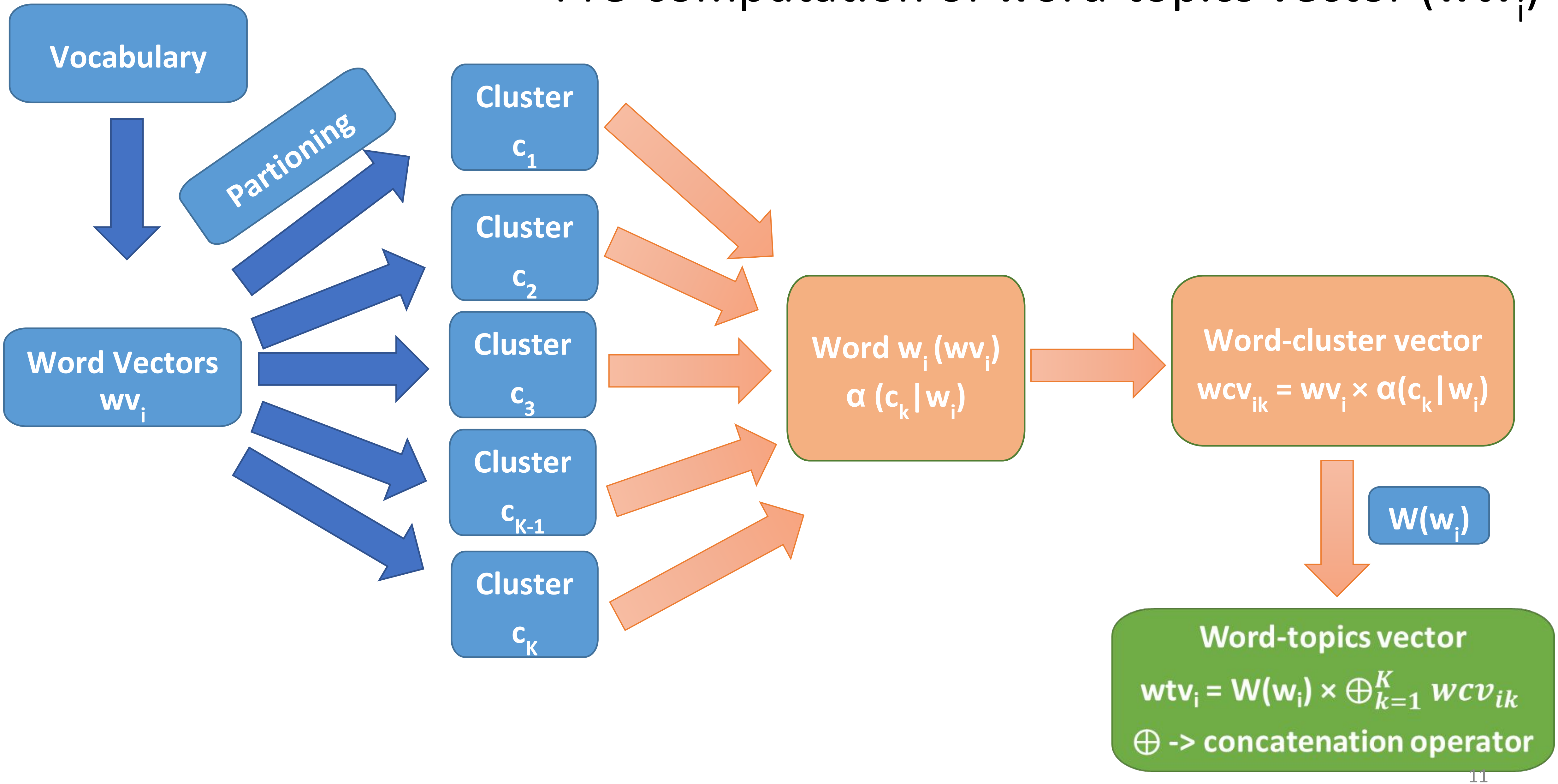
WEIGHTED PARTITION AVERAGING

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

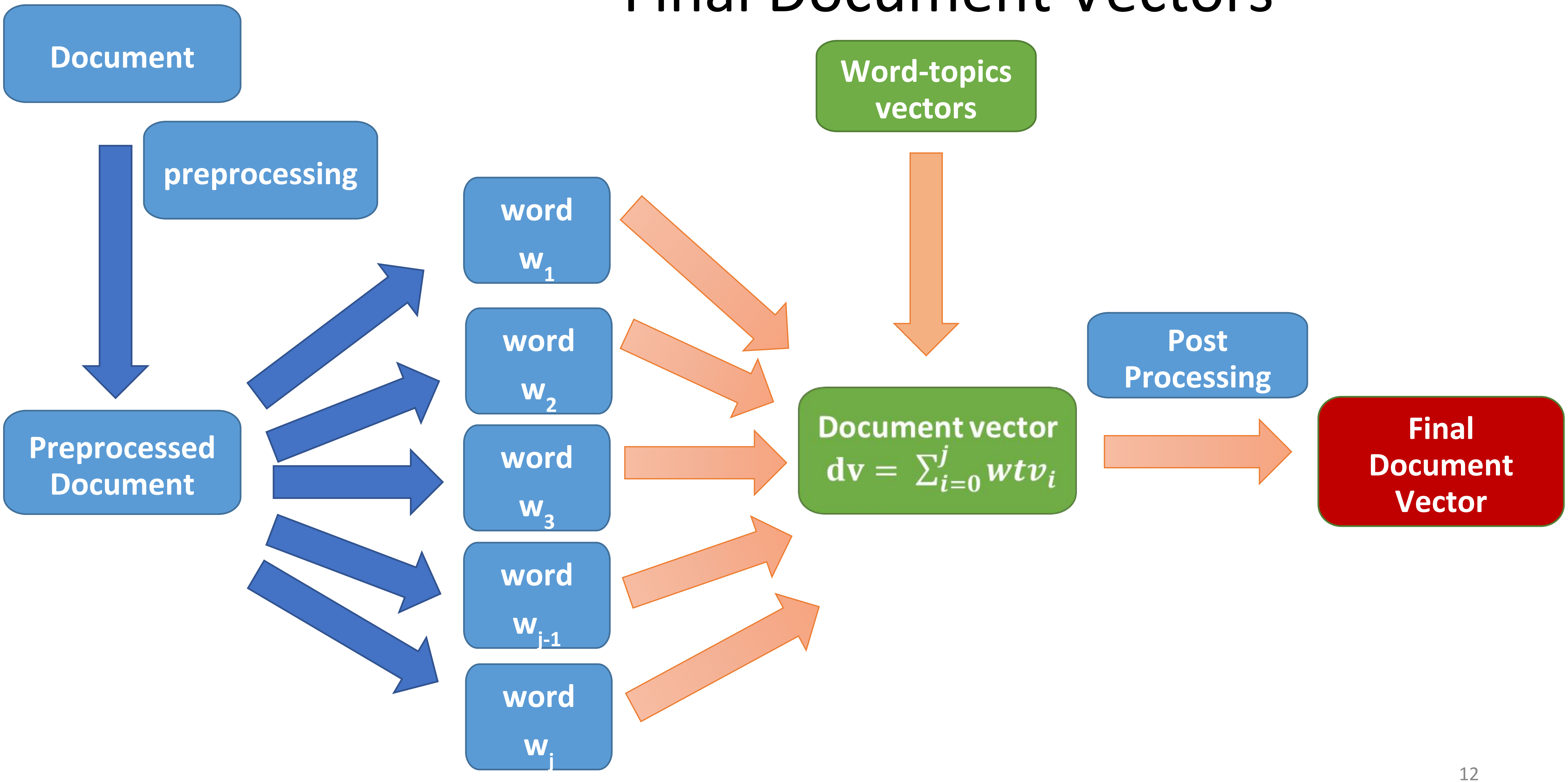
Proposed Algorithms (SCDV, P-SIF)

- Obtain word vectors for vocabulary words
- Partition vocabulary words using corresponding word vectors
  - K-Means
  - GMM
  - Sparse Dictionary
- For a document, do following
  - Weighted average intra partition
  - Concatenate averages inter partitions
- Post Processing Step
  - Hard Thresholding
  - Common Component Removal

Pre-computation of word-topics vector (wtv_i)



Final Document Vectors



Similar to simple weighted averaging model
we average **word topic vectors** instead of **word vectors**

Nice Connection

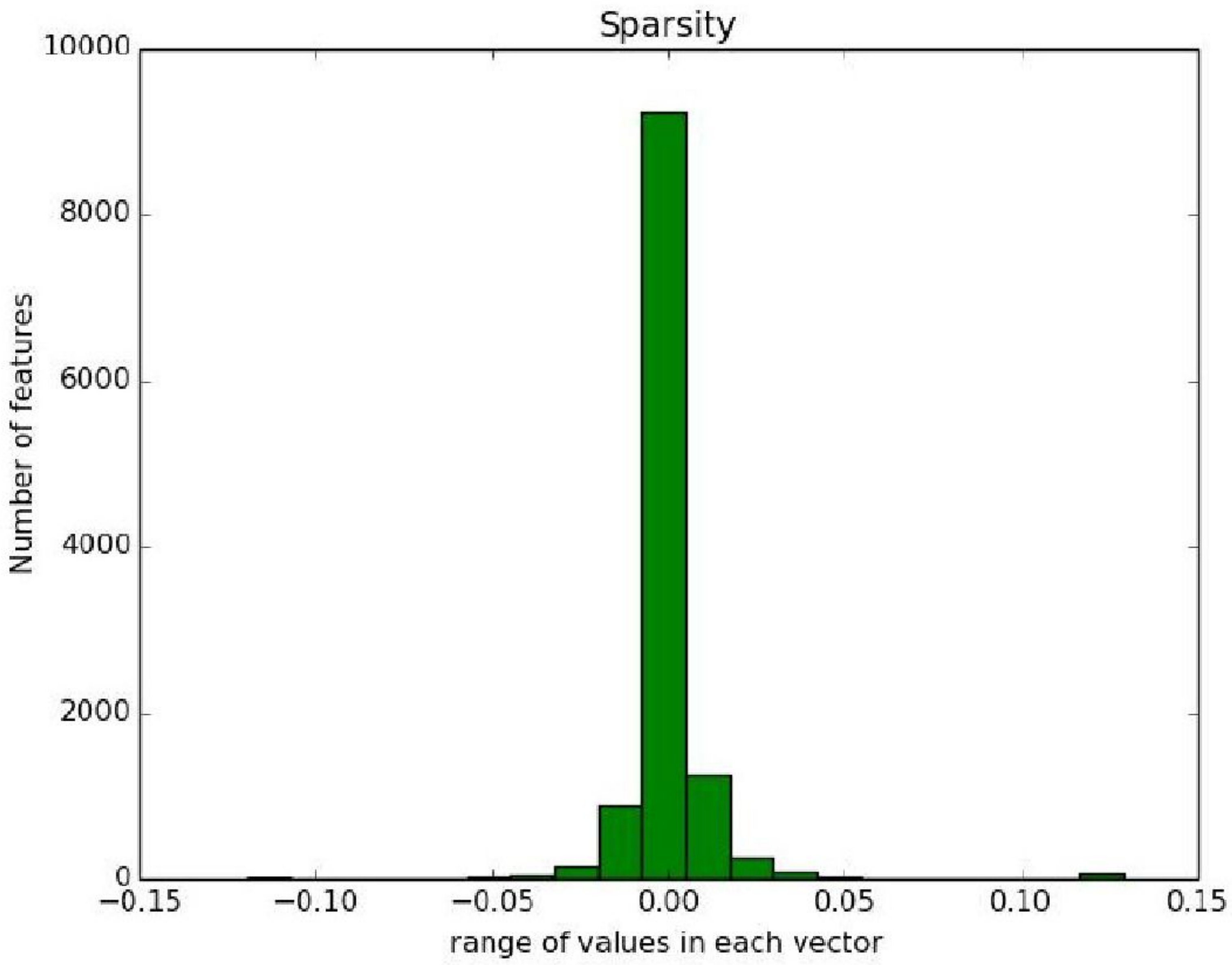
Several Partitioning Approaches

Name	Partition Type	Properties	Method
K-Means	Hard Clustering	Polysemic Words ☹, Vectors Sparsity ☺ , Partition Diversity ☹, Pre-Computation ☹	BOWV
GMM	Fuzzy Clustering	Polysemic Words ☺ , Vectors Sparsity☹, Partition Diversity☹, Manual Vector Sparsity (Hard Thresholding) ☺ , Pre-Computation ☺	SCDV
K-SVD	Sparse Dictionary Learning	Polysemic Words ☺ , Vectors Sparsity ☺ , Partition Diversity ☺ , Pre-Computation ☺	P-SIF

Weighting Algorithms

Technique	Operation	Method
Inverse document frequency	Concatenation	BOWV
Inverse document frequency	Multiplication	SCDV
Smooth Inverse frequency	Multiplication	P-SIF

Manual Sparsity by Hard Thresholding (SCDV)



Fuzzy vs Hard clustering? (SCDV context sensitive learning)

Word	Cluster Words	$P(C_i \mid W_j)$
Subject:1	Physics, chemistry, maths, science	0.27
Subject:2	Mail, letter, email, gmail	0.72
Interest:1	Information, enthusiasm, question	0.65
Interest:2	Bank, market, finance, investment	0.32
Break:1	Vacation, holiday, trip, spring	0.52
Break:2	Encryption, cipher, security, privacy	0.22
Break:2	If, elseif, endif, loop, continue	0.23
Unit:1	Calculation, distance, mass, length	0.25
Unit:2	Electronics, KWH, digital, signal	0.69

‘ similar results Arora et al., 2017

Topic Modelling using GMM

GMM	LTSG	LDA
-85.23	-92.33	-108.72

Topic Image			Topic Health			Topic Mail		
GMM	LTSG	LDA	GMM	LTSG	LDA	GMM	LTSG	LDA
file	image	image	heath	stimulation	doctor	ftp	anonymous	list
bit	jpeg	file	study	diseases	disease	mail	faq	mail
image	gif	color	medical	disease	coupons	internet	send	information
files	format	gif	drug	toxin	treatment	phone	ftp	internet
color	file	jpeg	test	toxic	pain	email	mailing	send
format	files	file	drugs	newsletter	medical	send	server	posting
images	convert	format	studies	staff	day	opinions	mail	email
jpeg	color	bit	disease	volume	microorganism	fax	alt	group
gif	formats	images	education	heaths	medicine	address	archive	news
program	images	quality	age	aids	body	box	email	anonymous
-67.16	-75.66	-88.79	-66.91	-96.98	-100.39	-77.47	-78.23	-95.47

Textual Classification

- Multi-Class Classification
  - 20 NewsGroup – 20 classes, Equal Sampling, 200-300 words documents

- Multi-Label Classification
  - Reuters - ~5000 labels, Unequal Sampling, 400-500 words documents

Multi-Class Classification – 20NewsGroup Dataset

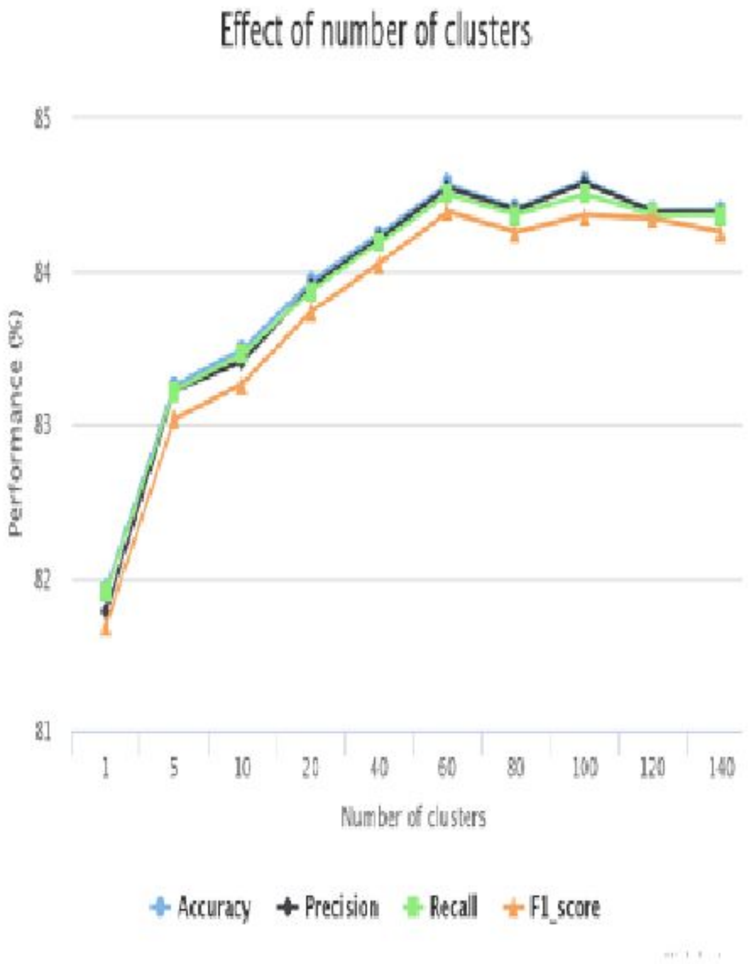
Model	Accuracy	Precision	Recall	F1-Score
P-SIF (Doc2VecC)	**86.0**	**86.1**	**86.1**	**86.0**
P-SIF	**85.4**	**85.5**	**85.4**	**85.2**
SCDV	**84.6**	**84.6**	**84.5**	**84.6**
BoE	83.1	83.1	83.1	83.1
BoWV	81.6	81.1	81.1	80.9
NTSG-1	82.6	82.5	81.9	81.2
LTSG	82.8	82.4	81.8	81.8
TWE-1	81.5	81.2	80.6	80.6
PV-DBoW	75.4	74.9	74.3	74.3
PV-DM	72.4	72.1	71.5	71.5

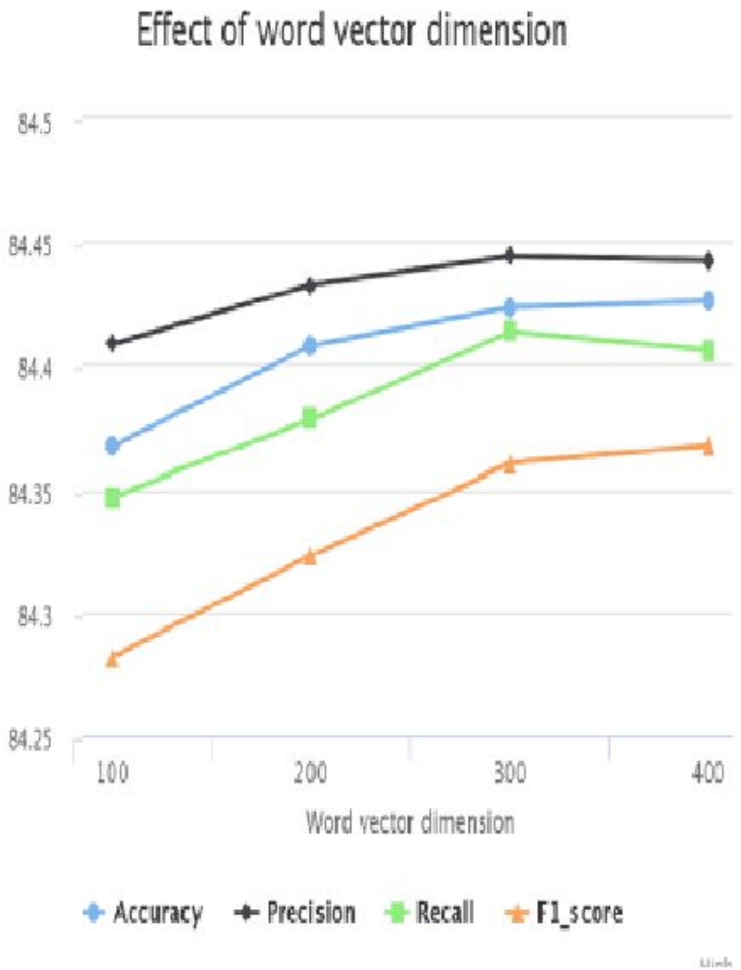
Time (sec)	BOWV	TWE-1	SCDV
Doc2Vec Formation	1250	700	**160**
Total Training	1320	740	**200**
Total Prediction	780	120	**25**

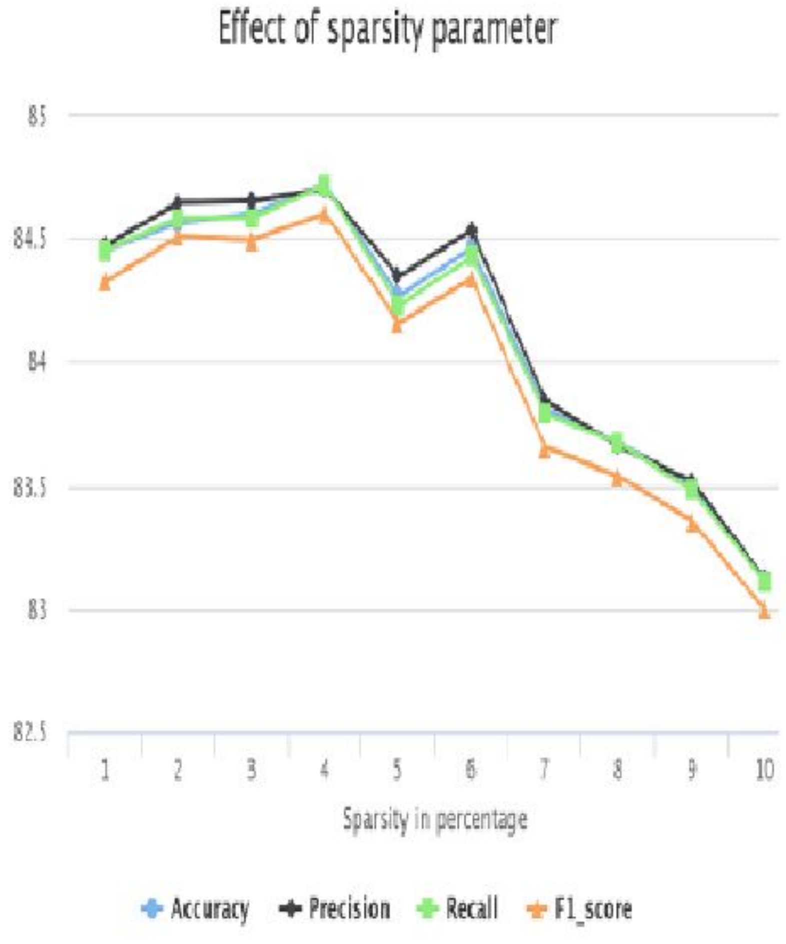
Class performance on 20 NewsGroup

	BoW		SCDV		P-SIF		P-SIF(Doc2VecC)	
Class Name	**Pre.**	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**	**Pre.**	**Rec.**
alt.atheism	67.8	72.1	80.2	79.5	**83.3**	**80.2**	83.0	79.9
comp.graphics	67.1	73.5	75.3	77.4	76.6	78.1	**76.8**	**79.2**
comp.os.ms-windows.misc	77.1	66.5	**78.6**	77.2	76.3	77.7	77.2	**78.2**
comp.sys.ibm.pc.hardware	62.8	72.4	**75.6**	73.5	73.4	**74.5**	71.1	74.2
comp.sys.mac.hardware	77.4	78.2	83.4	85.5	87.1	84.4	**87.5**	**87.5**
comp.windows.x	83.2	73.2	87.6	**78.6**	**89.3**	78.0	88.8	78.5
misc.forsale	81.3	88.2	81.4	85.9	**82.7**	**88.0**	82.4	86.4
rec.autos	80.7	82.8	91.2	90.6	**93.0**	90.1	92.8	**90.7**
rec.motorcycles	92.3	87.9	95.4	95.7	93.6	95.5	**97.0**	**96.5**
rec.sport.baseball	89.8	89.2	93.2	94.7	93.3	95.2	**95.2**	**95.7**
rec.sport.hockey	93.3	93.7	96.3	**99.2**	95.6	98.5	**96.8**	98.8
sci.crypt	92.2	86.1	92.5	**94.7**	89.8	93.2	**93.4**	96.7
sci.electronics	70.9	73.3	74.6	74.9	**79.6**	78.6	78.0	**79.3**
sci.med	79.3	81.3	91.3	88.4	91.9	88.6	**92.7**	**89.9**
sci.space	90.2	88.3	88.5	93.8	89.4	94.0	**90.7**	**94.4**
soc.religion.christian	77.3	87.9	83.3	92.3	84.0	**94.3**	**86.0**	92.5
talk.politics.guns	71.7	85.7	72.7	90.6	73.1	**91.2**	**77.3**	89.8
talk.politics.mideast	91.7	76.9	96.2	**95.4**	97.0	94.5	**97.5**	94.2
talk.politics.misc	71.7	56.5	80.9	59.7	81.0	59.0	**82.0**	**62.0**
talk.religion.misc	63.2	55.4	**73.5**	57.2	72.2	59.0	67.4	**62.4**

Effect of Hyperparameters (SCDV)







Multi-Label Classification - Reuters Dataset

Model	Prec@1 nDCG@1	Prec@5	nDCG@5	Coverage	LRAPS	F1-Score
P-SIF(Doc2VecC)	**94.92**	**37.98**	**50.40**	**6.03**	**93.95**	**82.87**
P-SIF	**94.77**	**37.33**	**49.97**	**6.24**	**93.72**	**82.41**
SCDV	**94.20**	**36.98**	**49.55**	**93.52**	**93.30**	**81.75**
BoWV	92.90	36.14	48.55	91.84	91.46	79.16
TWE-1	90.91	35.49	47.54	91.84	90.97	79.16
PV-DBoW	88.78	34.51	46.42	88.72	87.43	73.68
PV-DM	87.54	33.24	44.21	86.85	86.21	70.24

Mean average precision (MAP) on Information Retrieval Datasets

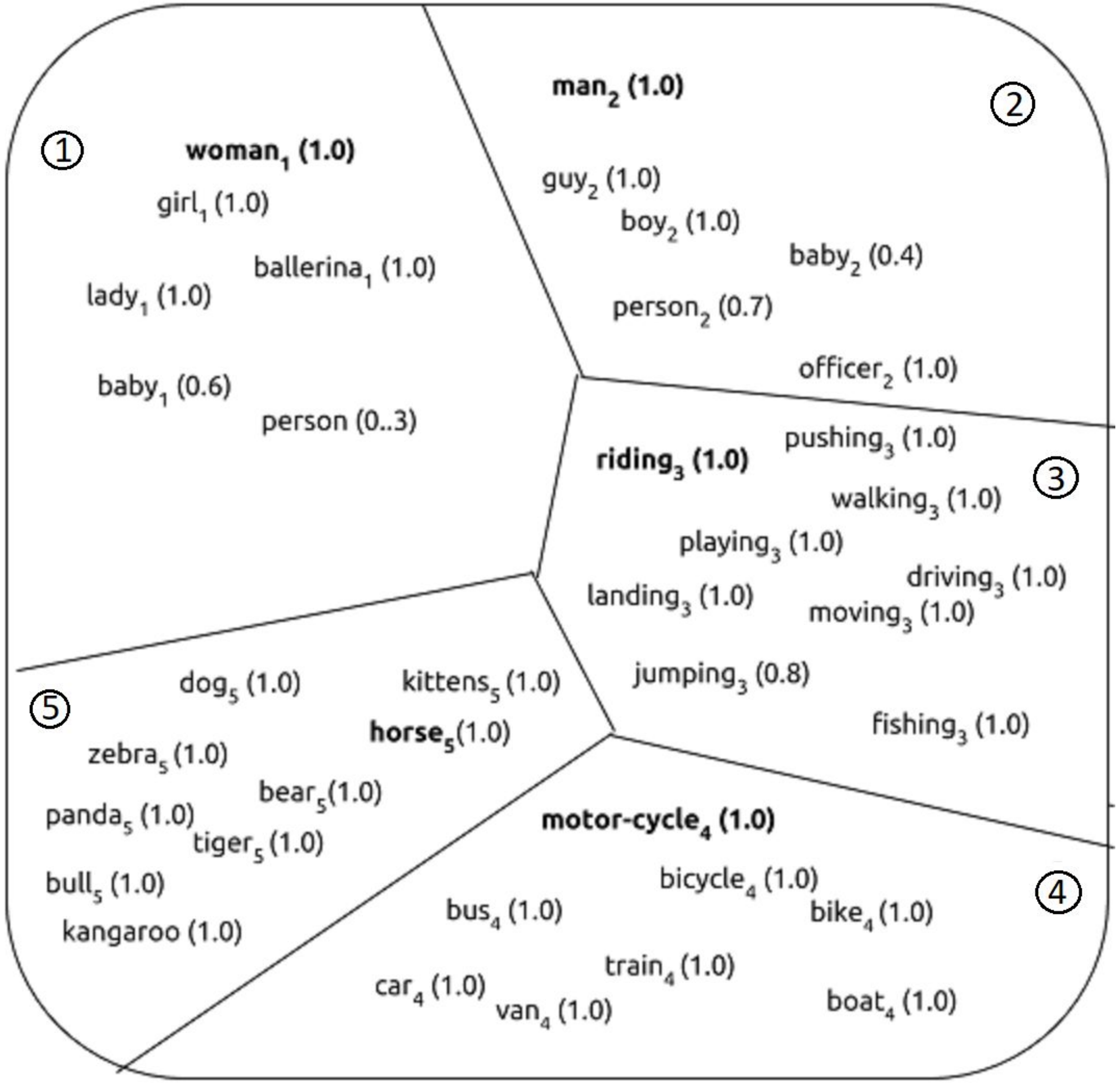
DataSet	LM	LM + SCDV	MB	MB + SCDV
AP	0.2742	**0.2856**	0.3283	**0.3395**
SJM	0.2052	**0.2105**	0.2341	**0.2409**
WSJ	0.2618	**0.2705**	0.3027	**0.3126**
Robust04	0.2516	**0.2684**	0.2819	**0.2933**

Comparison with WMD and WME

Dataset	Bbcsport	Twitter	Ohsumed	Classic	Reuters	Amazon	20NewsGroup	Recipe-L
BOW	79.4 ± 1.2	56.4 ± 0.4	38.9	64.0 ± 0.5	86.1	71.5 ± 0.5	42.2	-
TF-IDF	78.5 ± 2.8	66.8 ± 0.9	37.3	65.0 ± 1.8	70.9	58.5 ± 1.2	45.6	-
BM25	83.1F1.5	57.3 ± 7.8	33.8	59.4 ± 2.7	67.2	41.2 ± 2.6	44.1	-
LSI	95.7 ± 0.6	68.3 ± 0.7	55.8	93.3 ± 0.4	93.7	90.7 ± 0.4	71.1	-
LDA	93.6 ± 0.7	66.2 ± 0.7	49	95.0 ± 0.3	93.1	88.2 ± 0.6	68.5	-
mSDA	91.6 ± 0.8	67.7 ± 0.7	50.7	93.1 ± 0.4	91.9	82.9 ± 0.4	60.5	-
SIF(GloVe)	97.3 ± 1.2	57.8 ± 2.5	**67.1**	92.7 ± 0.9	87.6	94.1 ± 0.2	72.3	71.1 ± 0.5
Word2Vec +nbow	97.3 ± 0.9	72.0 ± 1.5	63	95.2 ± 0.4	96.9	94.0 ± 0.5	71.7	74.9 ± 0.5
Word2Vec +tf-idf	96.9 ± 1.1	71.9 ± 0.7	60.6	93.9 ± 0.4	95.9	92.2 ± 0.4	70.2	73.1 ± 0.6
PV-DBOW	97.2 ± 0.7	67.8 ± 0.4	55.9	97.0 ± 0.3	96.3	89.2 ± 0.3	71	73.1 ± 0.5
PV-DM	97.9 ± 1.3	67.3 ± 0.3	59.8	96.5 ± 0.7	94.9	88.6 ± 0.4	74	71.1 ± 0.4
Doc2VecC	90.5 ± 1.7	71.0 ± 0.4	63.4	96.6 ± 0.4	96.5	91.2 ± 0.5	78.2	76.1 ± 0.4
Doc2VecC (Train)	89.2 ± 1.4	69.8 ± 0.9	59.6	96.2 ± 0.5	96	89.5 ± 0.4	72.9	75.6 ± 0.4
KNN-WMD	95.4 ± 1.2	71.3 ± 0.6	55.5	97.2 ± 0.1	96.5	92.6 ± 0.3	73.2	71.4 ± 0.5
WME(SR)	95.5 ± 0.7	72.5 ± 0.5	55.8	96.6 ± 0.2	96	92.7 ± 0.3	72.9	72.5 ± 0.4
WME(LR)	98.2 ± 0.6	$\mathbf{74.5 \pm 0.5}$	64.5	97.1 ± 0.4	97.2	94.3 ± 0.4	78.3	$\mathbf{79.2 \pm 0.3}$
P-SIF	$\mathbf{99.05 \pm 0.9}$	73.39 ± 0.9	**67.1**	96.95 ± 0.5	**97.67**	94.17 ± 0.3	**79.15**	78.24 ± 0.3
P-SIF (Doc2VecC)	$\mathbf{99.68 \pm 0.9}$	72.39 ± 0.9	**67.1**	$\mathbf{97.7 \pm 0.5}$	**97.62**	$\mathbf{94.83 \pm 0.3}$	**86.31**	77.61 ± 0.3

Semantic Textual Similarity

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



	Document 1 (d_n^1)
Doc	A man is riding a motorcycle
SIF	$\vec{v}_{\text{man}_2} + \vec{v}_{\text{riding}_3} + \vec{v}_{\text{motorcycle}_4}$
P-SIF	$\vec{v}_{\text{zero}_1} \oplus \vec{v}_{\text{man}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{motorcycle}_4} \oplus \vec{v}_{\text{zero}_5}$

	Document 2 (d_n^2)
Doc	A woman is riding a horse
SIF	$\vec{v}_{\text{woman}_1} + \vec{v}_{\text{riding}_3} + \vec{v}_{\text{horse}_5}$
P-SIF	$\vec{v}_{\text{women}_1} \oplus \vec{v}_{\text{zero}_2} \oplus \vec{v}_{\text{riding}_3} \oplus \vec{v}_{\text{zero}_4} \oplus \vec{v}_{\text{horse}_5}$

SIMILARITY SCORES		
Ground Truth	SIF	P-SIF
0.15	0.57	0.16

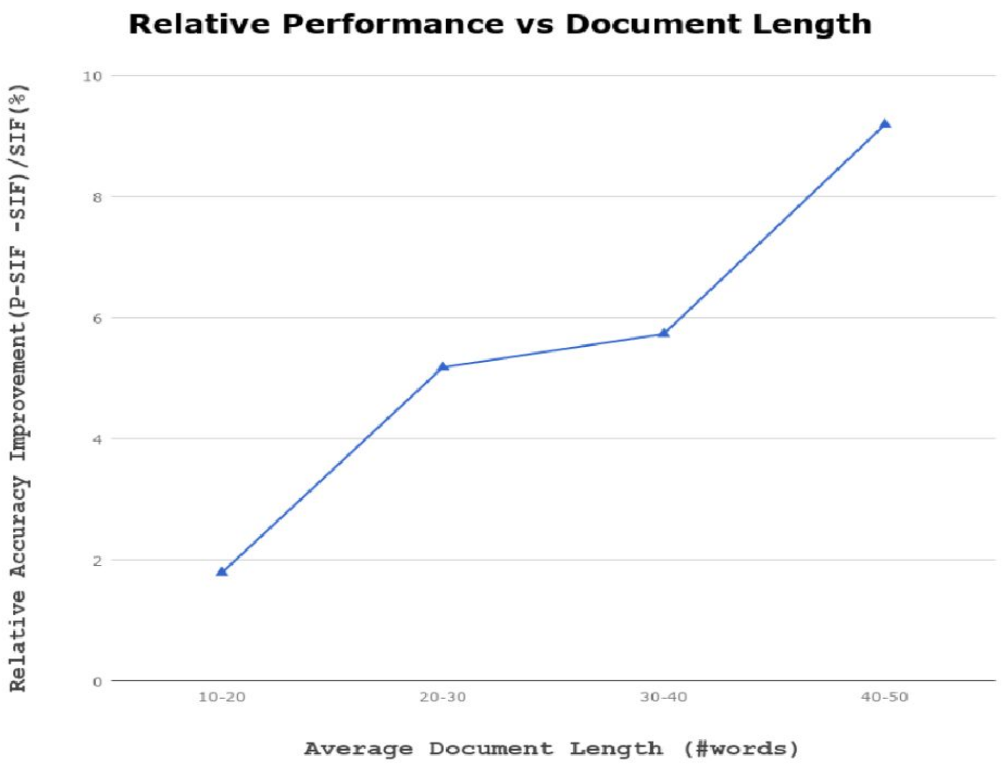
Results (Pearson r X 100) on Semantic Textual Similarity Task

Supervised or Not	Supervised								UnSupervised			Semi Supervised			P-SIF
Tasks	**PP**	**PP -Proj**	**DAN**	**RNN**	**iRNN**	**LSTM (no)**	**LSTM (o.g.)**	**GRAN**	**ST**	**avg Glove**	**tfidf Glove**	**avg -PSL**	**Glove +WR**	**PSL +WR**	**P-SIF +PSL**
STS12	58.7	60.0	56.0	48.1	58.4	51.0	46.4	62.5	30.8	52.5	58.7	52.8	56.2	59.5	**65.7**
STS13	55.8	56.8	54.2	44.7	56.7	45.2	41.5	63.4	24.8	42.3	52.1	46.4	56.6	61.8	**64.0**
STS14	70.9	71.3	69.5	57.7	70.9	59.8	51.5	**75.9**	31.4	54.2	63.8	59.5	68.5	73.5	74.8
STS15	75.8	74.8	72.7	57.2	75.6	63.9	56.0	**77.7**	31.0	52.7	60.6	60.0	71.7	76.3	77.3
SICK14	71.6	71.6	70.7	61.2	71.2	63.9	59.0	72.9	49.8	65.9	69.4	66.4	72.2	72.9	**73.4**
Twitter15	52.9	52.8	53.7	45.1	52.9	47.6	36.1	50.2	24.7	30.3	33.8	36.3	48.0	49.0	**54.9**

Results (Pearson r X 100) on Semantic Textual Similarity Task (16)

Tasks	Skip Thoughts	LSTM	Tree LSTM	Sent2Vec	Doc2Vec	avg Glove	tfidf Glove	avg PSL	tfidf PSL	Glove +WR	PSL +WR	P-SIF +PSL
STS16	51.4	64.9	64.0	73.7	69.4	47.2	51.1	63.3	66.9	72.4	72.5	**73.7**

Relative Performance (P-SIF –SIF / SIF %) Improvement on 26 STS



Theoretical Justification

- We showed connections of P-SIF with generative random-walk based latent variable models (Arora et. al. 2016a)
- Total number of topics in entire corpus (K) and can be determine by sparse dictionary learning (Arora et. al. 2016b)
- The context vector does not change significantly much while words are generated from random walk except topic change
- The partition function remain same in all directions for only words coming from a same context
- Taylor expansion followed by Maximum Likelihood Estimation over the distribution give the required context vector.
- Concatenation of context vector give the required document embedding.

Kernel Connection of embeddings

$$K^1(D_A, D_B) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle \vec{wv}_{w_i^A} \cdot \vec{wv}_{w_j^B}\rangle$$

- word vector averaging

$$K^2(D_A, D_B) = \frac{1}{nm}\langle \sum_{i=1}^{n} w\vec{t}v_{w_i^A} \cdot \sum_{j=1}^{m} w\vec{t}v_{w_j^B}\rangle$$

- Our P-SIF model

$$K^2(D_A, D_B) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle \vec{wv}_{w_i^A} \cdot \vec{wv}_{w_j^B}\rangle \times \langle \vec{tv}_{w_i^A} \cdot \vec{tv}_{w_j^B}\rangle$$

$$K^3(D_A, D_B) = \frac{1}{n}\sum_{i=1}^{n}\max_j\langle \vec{wv}_{w_i^A} \cdot \vec{wv}_{w_j^B}\rangle$$

$$K^5(D_A, D_B) = K^3(D_A, D_B) + K^4(D_A, D_B)$$

- Relax word mover distance
- Word mover distance

Summary

- Novel simple **unsupervised** technique to form **compositional** document vectors
  - Capture **distinctiveness** of words
  - Capture **semantics** of words
  - Represent **Sparse** & **Higher Dimension**
  - **Simple** and **Efficient**
- Perform SoTA on standard multi-class, multi-label classification, semantic textual similarity and information retrieval tasks.
- GMM clustering over words vectors can be used for context sensitive learning and topic modelling.
- Sparse dictionary produce diverse clusters, which reduces the size of the word topic vectors.

Future Directions

✓ Using multi sense embedding based on context in use instead of skip gram embeddings

✓ One can project the sparse word topic vector into a continuous low dimensional manifold, useful in downstream tasks especially deep learning

✓ Instead of using unsupervised weighting over word-topic vectors, one can learn weights in a supervise task

✗ Providing a more significant theoretical justification of embedding

✗ How we can take ordering into consideration e.g. LSTM along with partitioning

✗ Joint partitioning and classification (single step process)

Acknowledgement

I thanks the following people from the bottom of my heart

- Anonymous reviewer #2 of ICLR 2019 for important suggestions
- Prof. Vivek Srikumar for valuable suggestions especially in kernel derivations
- Prof. Aditya Bhaskara for discussion over the proofs
- Prof. Piyush Rai (IIT-K) for providing useful writing suggestions
- Pegah Nokhiz for reviewing the paper and slides at short notice
- Prof. Harish Karnick(IIT-K), Dheeraj Mekala (IIT-K), Bhargavi Paranjape (CMU)

Thanks for Listening

Questions?

Don’t Hesitate

email: *keviv9@gmail.com*

P-SIF Algorithm (Sparse Dictionary)

Algorithm 1: Main Algorithm

Data: Documents $\{D_n : D_n \in D\}$, Word embeddings $\{w\vec{v}_w : w \in V\}$, a set of sentences D, parameter a and estimated probabilities $\{p(w) : w \in V\}$ of the words, a sparsity parameter k, and an upper bound m.

Result: Document vectors $\{v\vec{D}_n : D_n \in D\}$

```
/* Dictionary learning for word-vectors */
```

for *each word w in V* **do**

  $wv_w = \sum_{j=1}^{m} \alpha_{w,j} A_j + \eta_w$;

end

```
/* Word topic-vector formation */
```

for *each word w in V* **do**

  for *each coefficients $\alpha_{w,j}$ of word w* **do**

    $wc\vec{v}_{ik} \leftarrow w\vec{v}_i \times \alpha_{w,j}$;

  end

  $wt\vec{v}_i \leftarrow \bigoplus_{k=1}^{K} wc\vec{v}_{ik}$;

  /* $\bigoplus$ is concatenation */

end

```
/* SIF reweighed document vector embedding */
```

for *each document D_n in D* **do**

  $v\vec{D}_n \leftarrow \sum_{w \in D_n} \frac{a}{a+p(w)} wt\vec{v}_w$;

end

Form a matrix X whose columns are $\{v\vec{D}_n : D_n \in D\}$, and let u be the first singular vector;

for *each document $D_n \in D$* **do**

  $v\vec{D}_n \leftarrow v\vec{D}_n$ - $uu^T v_{D_n}$;

end

Results (Pearson r X 100) on Semantic Textual Similarity Task

Supervised or not	Supervised								UnSupervised			Semi Supervised			P-SIF
Tasks	PP	PP -proj	DAN	RNN	iRNN	LSTM (no)	LSTM (o.g.)	GRAN	ST	avg -Glove	tfidf -Glove	avg -PSL	Glove +WR	PSL +WR	P-SIF +PSL
MSRpar	42.6	43.7	40.3	18.6	43.4	16.1	9.3	47.7	16.8	47.7	50.3	41.6	35.6	43.3	**52.4**
MSRvid	74.5	74.0	70.0	66.5	73.4	71.3	71.3	85.2	41.7	63.9	77.9	60.0	83.8	84.1	**85.6**
SMT-eur	47.3	49.4	43.8	40.9	47.1	41.8	44.3	49.3	35.2	46.0	54.7	42.4	49.9	44.8	**58.7**
OnWN	70.6	70.1	65.9	63.1	70.1	65.2	56.4	71.5	29.7	55.1	64.7	63.0	66.2	71.8	**72.2**
SMT-news	58.4	**62.8**	60.0	51.3	58.1	60.8	51.0	58.7	30.8	49.6	45.7	57.0	45.6	53.6	59.5
STS12	58.7	60.0	56.0	48.1	58.4	51.0	46.4	62.5	30.8	52.5	58.7	52.8	56.2	59.5	**65.7**
headline	72.4	72.6	71.2	59.5	72.8	57.4	48.5	**76.1**	34.6	63.8	69.2	68.8	69.2	74.1	75.7
OnWN	67.7	68.0	64.1	54.6	69.4	68.5	50.4	81.4	10.0	49.0	72.9	48.0	82.8	82.0	**84.4**
FNWN	43.9	46.8	43.1	30.9	45.3	24.7	38.4	**55.6**	30.4	34.2	36.6	37.9	39.4	52.4	54.8
SMT	39.2	39.8	38.3	33.8	39.4	30.1	28.8	40.3	24.3	22.3	29.6	31.0	37.9	38.5	**41.0**
STS13	55.8	56.8	54.2	44.7	56.7	45.2	41.5	63.4	24.8	42.3	52.1	46.4	56.6	61.8	**64.0**
deft forum	48.7	51.1	49.0	41.5	49.0	44.2	46.1	**55.7**	12.9	27.1	37.5	37.2	41.2	51.4	53.2
deft news	73.1	72.2	71.7	53.7	72.4	52.8	39.1	**77.1**	23.5	68.0	68.7	67.0	69.4	72.6	75.2
headline	69.7	70.8	69.2	57.5	70.2	57.5	50.9	**72.8**	37.8	59.5	63.7	65.3	64.7	70.1	70.2
images	78.5	78.1	76.9	67.6	78.2	68.5	62.9	**85.8**	51.2	61.0	72.5	62.0	82.6	84.8	84.8
OnWN	78.8	79.5	75.7	67.7	78.8	76.9	61.7	85.1	23.3	58.4	75.2	61.1	82.8	84.5	**88.1**
tweet news	76.4	75.8	74.2	58.0	76.9	58.7	48.2	**78.7**	39.9	51.2	65.1	64.7	70.1	77.5	77.5
STS14	70.9	71.3	69.5	57.7	70.9	59.8	51.5	**75.8**	31.4	54.2	63.8	59.5	68.5	73.5	74.8
answers-forum	68.3	65.1	62.6	32.8	67.4	51.9	50.7	**73.1**	36.1	30.5	45.6	38.8	63.9	70.1	70.7
answers-student	78.2	77.8	78.1	64.7	78.2	71.5	55.7	72.9	33.0	63.0	63.9	69.2	70.4	75.9	**79.6**
belief	76.2	75.4	72.0	51.9	75.9	61.7	52.6	**78**	24.6	40.5	49.5	53.2	71.8	75.3	75.3
headline	74.8	75.2	73.5	65.3	75.1	64.0	56.6	**78.6**	43.6	61.8	70.9	69.0	70.7	75.9	76.8
images	81.4	80.3	77.5	71.4	81.1	70.4	64.2	**85.8**	17.7	67.5	72.9	69.9	81.5	84.1	84.1
STS15	75.8	74.8	72.7	57.2	75.6	63.9	56.0	**77.7**	31.0	52.7	60.6	60.0	71.7	76.3	77.3
SICK14	71.6	71.6	70.7	61.2	71.2	63.9	59.0	72.9	49.8	65.9	69.4	66.4	72.2	72.9	**73.4**
Twitter15	52.9	52.8	53.7	45.1	52.9	47.6	36.1	50.2	24.7	30.3	33.8	36.3	48.0	49.0	**54.9**

Results (Pearson r X 100) on Semantic Textual Similarity Task (16)

Tasks	Skip Thoughts	LSTM	Tree LSTM	Sent2Vec	Doc2Vec	avg Glove	tfidf Glove	avg PSL	tfidf PSL	Glove +WR	PSL +WR	P-SIF +PSL
STS16	51.4	64.9	64.0	73.7	69.4	47.2	51.1	63.3	66.9	72.4	72.5	**73.7**

Tasks	Skip thoughts	LSTM	Tree LSTM	Sent2Vec	Doc2Vec	Glove Avg	Glove tf-idf	PSL Avg	PSL tf-idf	Glove +WR	PSL +WR	P-SIF +PSL
headlines	51.019	75.7	74.08	75.06	69.16	49.66	52.76	70.86	72.24	72.86	74.48	**75.6**
plagiarism	66.708	71.73	67.62	80.06	80.6	59.84	61.48	77.96	80.06	79.46	79.74	**81.6**
post editing	69.947	72.31	70.65	82.85	82.85	59.89	62.34	80.41	81.45	82.03	82.05	**83.7**
answer answer	28.626	44.17	52.27	57.73	41.12	19.8	22.47	38.5	41.56	58.15	59.98	**60.2**
question question	40.459	60.69	55.26	73.03	**73.03**	46.84	56.58	48.69	59.1	69.36	66.41	67.2
STS16	51.4	64.9	64.0	**73.7**	69.4	47.2	51.1	63.3	66.9	72.4	72.5	**73.7**

Positive Qualitative Results (MSRvid)

sentence1	sentence2	GT	NGT	SIF_{sc}	$P\text{-}SIF_{sc}$
People are playing baseball .	The cricket player hit the ball .	0.5	0.1	0.2928	0.0973
A woman is carrying a boy .	A woman is carrying her baby .	2.333	0.4666	0.5743	0.4683
A man is riding a motorcycle .	A woman is riding a horse .	0.75	0.15	0.5655	0.157
A woman slices a lemon .	A man is talking into a microphone .	0	0	-0.1101	-0.0027
A man is hugging someone .	A man is taking a picture .	0.4	0.08	0.2021	0.0767
A woman is dancing .	A woman plays the clarinet .	0.8	0.16	0.3539	0.1653
A train is moving .	A man is doing yoga .	0	0	0.1674	-0.0051
Runners race around a track .	Runners compete in a race .	3.2	0.64	0.7653	0.6438
A man is driving a car .	A man is riding a horse .	1.2	0.24	0.3584	0.2443
A man is playing a guitar .	A woman is riding a horse .	0.5	0.1	-0.0208	0.0955
A man is riding on a horse .	A girl is riding a horse .	2.6	0.52	0.6933	0.5082
A woman is deboning a fish .	A man catches a fish .	1.25	0.25	0.4538	0.2336
A man is playing a guitar .	A man is eating pasta .	0.533	0.1066	-0.0158	0.0962
A woman is dancing .	A man is eating .	0.143	0.0286	-0.1001	0.0412
The ballerina is dancing .	A man is dancing .	1.75	0.35	0.512	0.3317
A woman plays the guitar .	A man sings and plays the guitar .	1.75	0.35	0.5036	0.3683
A girl is styling her hair .	A girl is brushing her hair .	2.5	0.5	0.7192	0.5303
A guy is playing hackysack	A man is playing a key-board .	1	0.2	0.3718	0.2268
A man is riding a bicycle .	A monkey is riding a bike .	2	0.4	0.6891	0.4614
A woman is swimming underwater .	A man is slicing some carrots .	0	0	-0.2158	-0.0562
A plane is landing .	A animated airplane is landing .	2.8	0.56	0.801	0.6338
The missile exploded .	A rocket exploded .	3.2	0.64	0.8157	0.6961
A woman is peeling a potato .	A woman is peeling an apple .	2	0.4	0.6938	0.5482
A woman is writing .	A woman is swimming .	0.5	0.1	0.3595	0.2334
A man is riding a bike .	A man is riding on a horse .	2	0.4	0.6781	0.564
A panda is climbing .	A man is climbing a rope .	1.6	0.32	0.4274	0.3131
A man is shooting a gun .	A man is spitting .	0	0	0.2348	0.1305

Negative Qualitative Results (MSRvid)

sentence1	sentence2	GT	NGT	SIF_{sc}	P-SIF_{sc}
takes off his sunglasses .	A boy is screaming .	0.5	0.1	0.1971	0.3944
The rhino grazed on the grass .	A rhino is grazing in a field .	4	0.8	0.7275	0.538
An animal is biting a persons finger .	A slow loris is biting a persons finger .	3	0.6	0.6018	0.7702
Animals are playing in water .	Two men are playing ping pong .	0	0	0.0706	0.2238
Someone is feeding a animal .	Someone is playing a piano .	0	0	-0.0037	0.1546
The lady sliced a tomatoe .	Someone is cutting a tomato .	4	0.8	0.693	0.5591
The lady peeled the potatoe .	A woman is peeling a potato .	4.75	0.95	0.7167	0.5925
A man is slicing something .	A man is slicing a bun .	3	0.6	0.5976	0.4814
A boy is crawling into a dog house .	A boy is playing a wooden flute .	0.75	0.15	0.1481	0.2674
A man and woman are talking .	A man and woman is eating .	1.6	0.32	0.3574	0.4711
A man is cutting a potato .	A woman plays an electric guitar .	0.083	0.0166	-0.1007	-0.2128
A person is cutting a meat .	A person riding a mechanical bull	0	0	0.0152	0.1242
A woman is playing the flute .	A man is playing the guitar .	1	0.2	0.1942	0.0876

Major References

- **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
- **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 embedding's"* In Proc AAAI, 2015
- **Lda2Vec** : Chris Moody *"Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec"*, arXiv:1605.02019
- **WMD** : Matt J. Kusner et al., *"From Word Embeddings To Document Distance*", In ICML 2015
- **WME** : Lingfei Wu, Ian E.H. Yen et. al., *"Word Mover's Embedding: From Word2Vec to Document Embedding",* In EMNLP 2018
- **weigh-AvgVec** : 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", arXiv preprint arXiv:1601.03764*
- ***Doc2vec** :* Quoc V Le and Tomas Mikolov. *"Distributed Representations of Sentences andDocuments"* In: ICML 2014