



Comparative Analysis of TF-IDF and TF-IDF Groups using Glove and GloveBERT



TASI-2122-103

- Alex Conro Manuel
- Angelina Naomi C Sinaga




Object

- Background
- Goals
- Dataset
- Framework
- Demo Project
- Evaluation
- Conclusion




Background

- 
- Belum ada penelitian sebelumnya yang melakukan perbandingan pembobotan TF-IDF dengan TF-IDF Group menggunakan GloVe dan GloVeBERT.



Goals

- 
- Membuat perbandingan pada TF-IDF dan TF-IDF Group dengan model GloVe dan GloVeBERT.
 - Melakukan pengelompokan dokumen menggunakan Mini-batch K-Means
 - Menghitung Cosine Similarity untuk memeriksa kesamaan dari satu dokumen terhadap dokumen lain
 - Melakukan eksperimen pada 2 dataset



Dataset



Dataset yang digunakan ada 2 yaitu

- Spam dataset
(<https://tinyurl.com/spamdataset>)
- BBC News dataset
(<https://tinyurl.com/bbcdataset>)



Framework

Glove

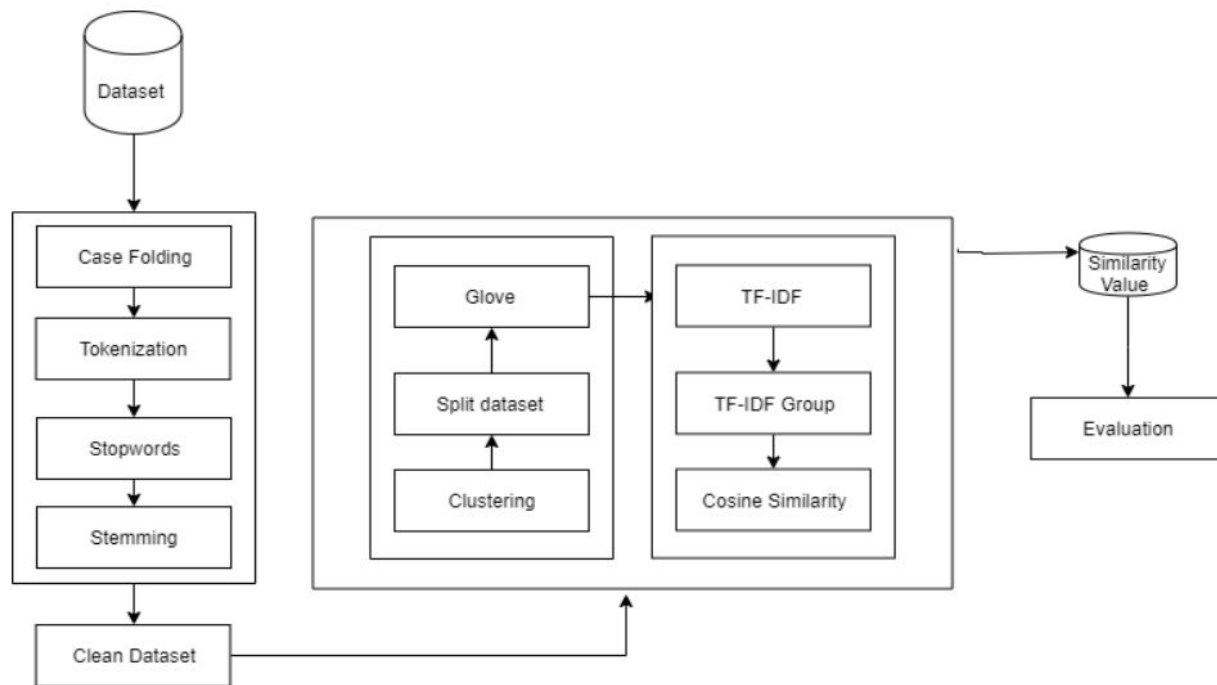


Fig. 1: Framework Glove

GloveBERT

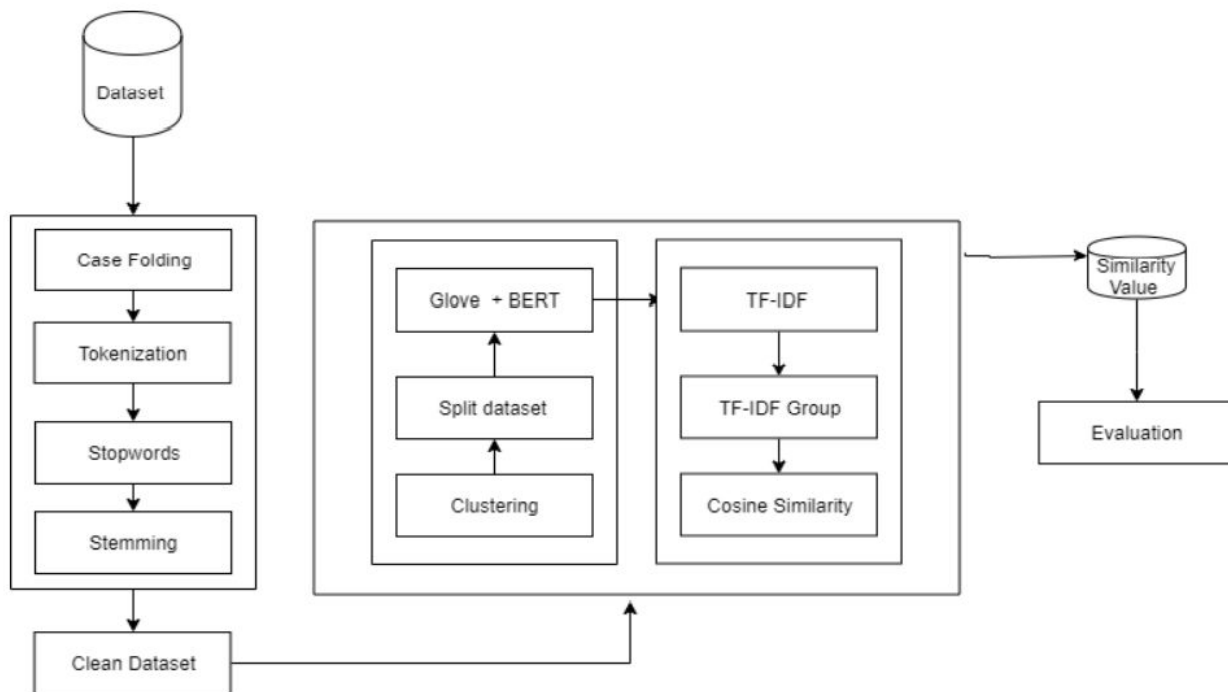


Fig. 2: Framework GloveBERT



Demo Project



Evaluation



Tabel Evaluasi Accuracy dan Cosine Similarity

Dataset	Method	Cosine Similarity	Accuracy
Spam	Glove	0.738	0.977
	GloveBERT	0.677	0.954
BBC News	Glove	0.774	0.058
	GloveBERT	0.736	0.0010



Evaluasi TF-IDF VS TF-IDF Group in Glove and GloveBERT

Glove in Spam Dataset

term	TF-IDF	TF-IDF-Group
free	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
entry	0.0003589375448671931	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788]
	0.0010768126346015793	[0.00215363 0.00215363 0.00215363 ... 0.00215363 0.00215363 0.00215363]
wkly	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
comp	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
win	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
fa	0.0003589375448671931	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788]
cup	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
final	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
tkts	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
may	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
.	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
text	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
receive	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
estion	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
(0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]

Glove in BBC News Dataset

term	TF-IDF	TF-IDF-Group
jeremy	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
bowen	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
	0.006482982171799027	[0.06482982 0.06482982 0.06482982 ... 0.06482982 0.06482982 0.06482982]
frontline	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
irpin	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
,	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
residents	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
came	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
russian	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
fire	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
trying	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
flee	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
.	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]

GloveBERT in Spam Dataset

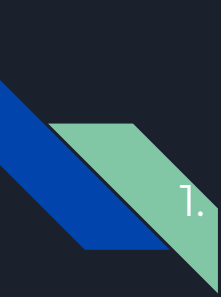
term	TF-IDF	TF-IDF-Group
free	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
entry	0.0003589375448671931 0.0010768126346015793	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788] [0.00215363 0.00215363 0.00215363 ... 0.00215363 0.00215363 0.00215363]
wkly	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
comp	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
win	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
fa	0.0003589375448671931	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788]
cup	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
final	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
tkts	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
may	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
.	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
text	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
receive	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
question	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
(0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]

GloveBERT in BBC News Dataset

term	TF-IDF	TF-IDF-Group
jeremy	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
bowen	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
	0.006482982171799027	[0.06482982 0.06482982 0.06482982 ... 0.06482982 0.06482982 0.06482982]
frontline	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
irpin	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
,	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
residents	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
came	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
russian	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
fire	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
trying	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
flee	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]
.	0.0008103727714748784	[0.00810373 0.00810373 0.00810373 ... 0.00810373 0.00810373 0.00810373]



Conclusion

- 
1. Pada perbandingan TF-IDF dengan TF-IDF Group nilai TF-IDF Group lebih tinggi dari TF-IDF pada pendekatan GloVe dan GloVeBERT untuk dataset Spam.
 2. Pada perbandingan TF-IDF dengan TF-IDF Group nilai TF-IDF memiliki nilai yang sama dengan TF-IDF Group pada pendekatan GloVe dan pendekatan GloVeBERT untuk dataset BBC News.
 3. Nilai Cosine Similarity pada kedua dataset cukup tinggi, hal ini menunjukkan bahwa satu dokumen memiliki kemiripan dengan yang lain pada kedua dataset.
 4. Menggabungkan GloVe dengan BERT memiliki efek penurunan baik pada nilai Cosine Similarity maupun pada nilai akurasi model.
 5. Berbeda dengan nilai akurasi pada dataset Spam pada kedua model, nilai akurasi dari model GloVe dan GloVeBERT pada dataset BBC News tergolong rendah.

Terima Kasih

