Comparative Analysis of TF-IDF and TF-IDF Groups using BERT and GloVe

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 BERT dan Glove.

Goals

1. Membuat perbandingan pada TF IDF dan TF-IDF Group dengan model BERT dan Glove.

2. Mengelompokkan dokumen menggunakan Mini batch dan menghitung Cosine Similarity.

3. Melihat performa pada dua set data, yaitu Spam dan BBC News untuk membandingkan hasil dari dua metode yang diterapkan.

Dataset

Dataset yang digunakan ada 2 yaitu

- Spam dataset (https://tinyurl.com/spamdataset)

- BBC News dataset (https://tinyurl.com/bbcdataset)

Framework

Glove

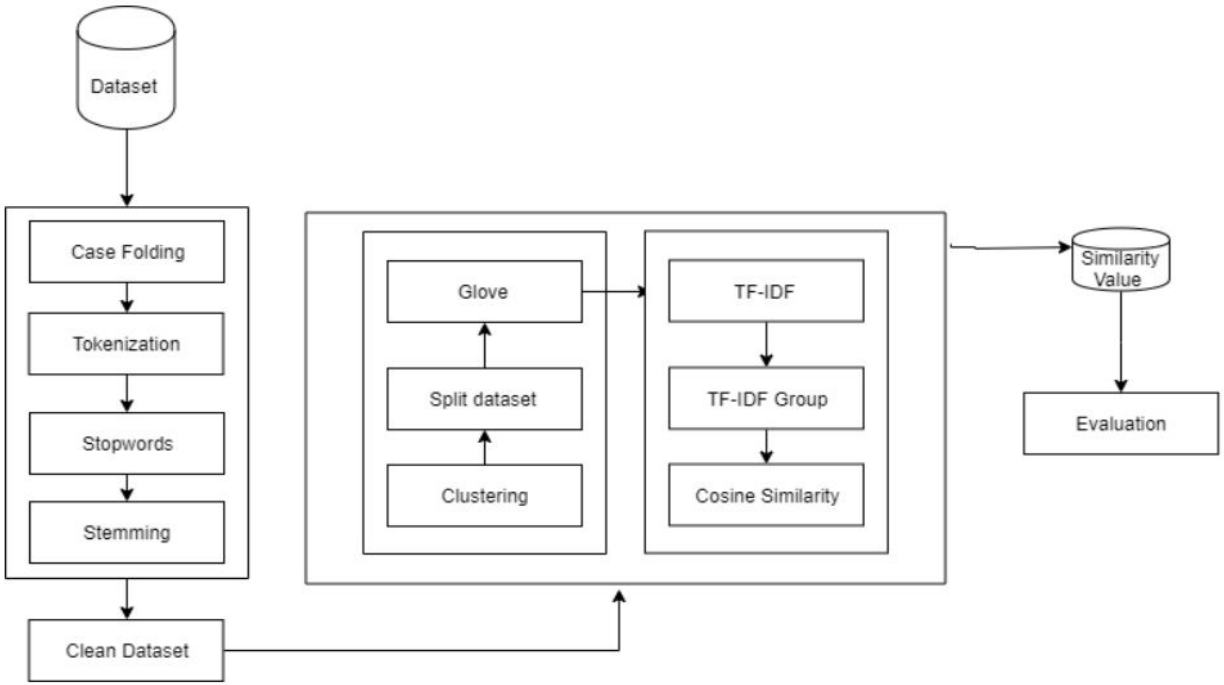

Fig. 1: Framework Glove

GloveBERT

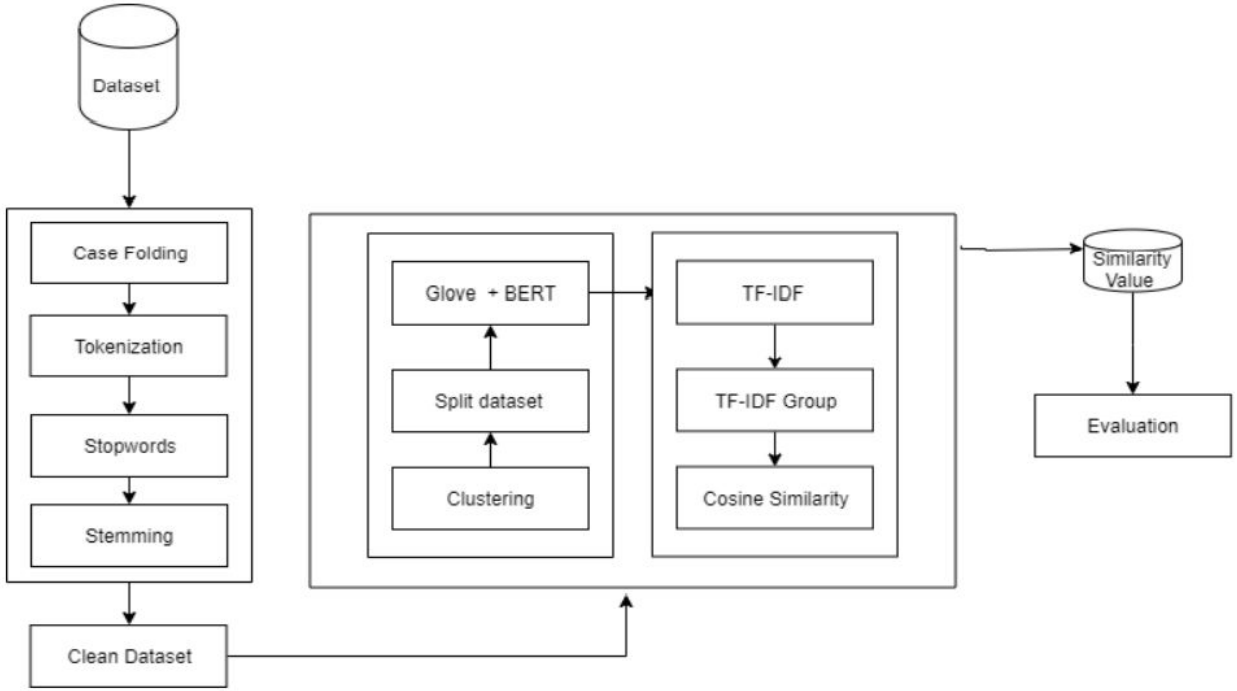

Fig. 2: Framework GloveBERT

Demo Project

Evaluation

Tabel Evaluasi

Dataset	Method	Cosine Similarity	Accuracy
Spam	Glove	0.710	0.977
	GloveBERT	0.677	0.954
BBC News	Glove	0.683	0.058
	GloveBERT	0.640	0.0010

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

Glove in Dataset 1

term	TF-IDF	TF-IDF-Group
9	0.13	[0.26 0.26 0.26 ... 0.26 0.26 0.26]
1	0.09	[0.18 0.18 0.18 ... 0.18 0.18 0.18]
0	0.11	[0.22 0.22 0.22 ... 0.22 0.22 0.22]
3	0.13	[0.26 0.26 0.26 ... 0.26 0.26 0.26]
7	0.07	[0.14 0.14 0.14 ... 0.14 0.14 0.14]
2	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]
5	0.09	[0.18 0.18 0.18 ... 0.18 0.18 0.18]
8	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]
4	0.08	[0.16 0.16 0.16 ... 0.16 0.16 0.16]
6	0.1	[0.2 0.2 0.2 ... 0.2 0.2 0.2]

Glove in Dataset 2

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 Dataset 1

term	TF-IDF	TF-IDF-Group
ok	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
lar	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
...	0.0003589375448671931	[0.00071788 0.00071788 0.00071788 ... 0.00071788 0.00071788 0.00071788]
joking	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
wif	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
u	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]
oni	0.00017946877243359656	[0.00035894 0.00035894 0.00035894 ... 0.00035894 0.00035894 0.00035894]

GloveBERT in Dataset 2

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

- Pada kedua dataset, nilai Cosine Similarity dan nilai akurasi cukup tinggi.
- Menggabungkan Glove dengan BERT memberikan efek penurunan nilai cosine similarity dan penurunan nilai akurasi pada model.
- Nilai TF-IDF pada Dataset 1 mengalami kenaikan saat dilakukan TF-IDF Group dengan metode GloVe dan GloveBERT.
- Nilai TF-IDF pada Dataset 2 memiliki nilai yang sama dengan TF-IDF Group dengan metode GloVe dan GloveBERT.

Terima Kasih