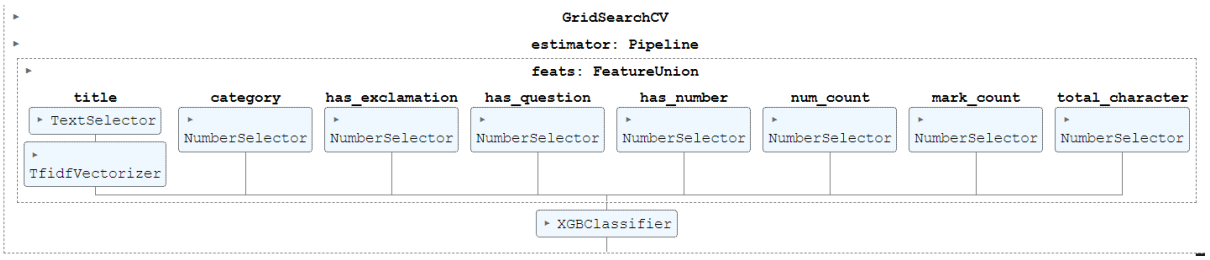
Catatan Model XGBoost, Random Forest, Naive Bayes

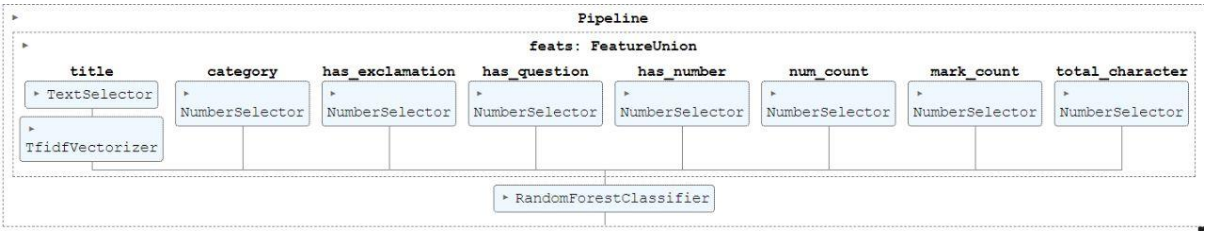
Hyperparameter terbaik untuk model *XGBoost* yang ditemukan peneliti adalah: random state=42, n estimators=210, colsample_bytree=0.65, subsample=1, learning_rate=0.1, max_depth=12, reg_lambda=1, seed=4. *Hyperparameter* ini didapatkan melalui *GridSearchCV*. Hal ini berbeda dengan model *Random Forest* dan *Naïve Bayes* kami yang menggunakan *baseline model*.

Alur Kerja Setiap Model

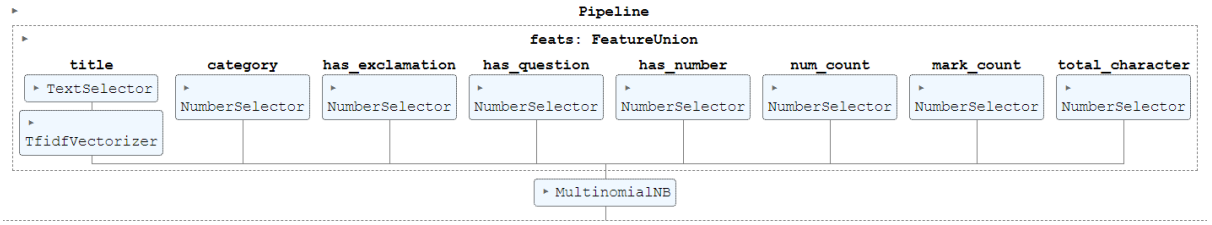
XGBoost



Random Forest



Naïve Bayes



Keterangan:
title: judul berita
category: kategori berita
has_exclamation: apakah judul mengandung tanda seru
has_question: apakah judul mengandung tanda tanya
num_count: jumlah angka muncul pada judul
mark_count: jumlah tanda baca pada judul
total_character: jumlah karakter pada judul

TABLE II
MODEL EVALUATION

Model	Title	
	Accuracy	Recall
XGBoost	0.7985	0.6653
Random Forest	0.7950	0.6720
Naive Bayes	0.7606	0.5532
Model	**Title + Content**	
	Accuracy	**Recall**
XGBoost	0.8094	0.7063
Random Forest	0.7342	0.5311
Naive Bayes	0.7249	0.5603
Model	**Augmented Title**	
	Accuracy	**Recall**
XGBoost	0.8001	0.6880
Random Forest	0.8014	0.6994
Naive Bayes	0.7608	0.6222
Model	**Augmented Title + Content**	
	Accuracy	**Recall**
XGBoost	0.7997	0.6879
Random Forest	0.7395	0.5378
Naive Bayes	0.7142	0.6509

Tabel berikut merupakan hasil skor *accuracy* dan *recall* yang didapatkan melalui *train* dan *test* dari data yang memiliki jumlah 14.000an baris. Kami merasa metode *labelling* data tersebut kurang kredibel karena hanya dilakukan oleh satu orang. Oleh karena itu, kami memutuskan untuk menggunakan data yang terdiri dari 8.000an baris dimana ketiga data tersebut dilakukan *labelling* dengan kesepakatan dari tiga orang, sehingga data tersebut menjadi lebih kredibel. Data tersebut pun menghasilkan skor *accuracy* dan *recall* yang lebih baik dibanding data yang memiliki 14.000an baris.