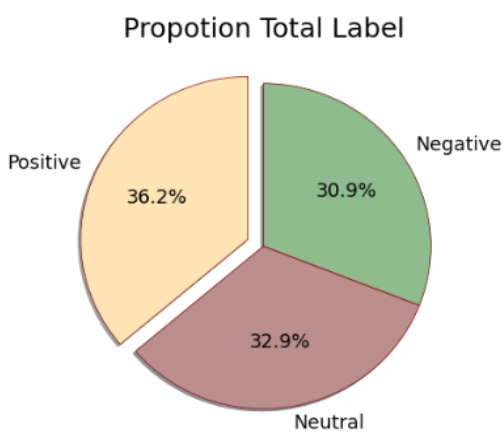


BEFORE PREPROCESSING

	Headlines	Time	Description
0	Jim Cramer: A better way to invest in the Covi...	7:51 PM ET Fri, 17 July 2020	"Mad Money" host Jim Cramer recommended buying...
1	Cramer's lightning round: I would own Teradyne	7:33 PM ET Fri, 17 July 2020	"Mad Money" host Jim Cramer rings the lightnin...
2	NaN	NaN	NaN
3	Cramer's week ahead: Big week for earnings, ev...	7:25 PM ET Fri, 17 July 2020	"We'll pay more for the earnings of the non-Co...
4	IQ Capital CEO Keith Bliss says tech and healt...	4:24 PM ET Fri, 17 July 2020	Keith Bliss, IQ Capital CEO, joins "Closing Be...
...
3075	Markets lack Christmas cheer	10:15 AM ET Tue, 26 Dec 2017	According to Kensho, here's how markets have f...
3076	Cramer Remix: The biggest mistake you can make...	11:12 AM ET Thu, 20 Sept 2018	Jim Cramer revealed his top rule when it comes...
3077	Cramer says owning too many stocks and too lit...	7:07 PM ET Fri, 22 Dec 2017	Jim Cramer broke down why owning fewer stocks ...
3078	Cramer: I helped investors through the 2010 fl...	7:07 PM ET Fri, 22 Dec 2017	Jim Cramer built on his "nobody ever made a di...
3079	Cramer: Never buy a stock all at once — you'll...	6:52 PM ET Fri, 22 Dec 2017	Jim Cramer doubled down on his key investing r...



Alur Pengerjaan

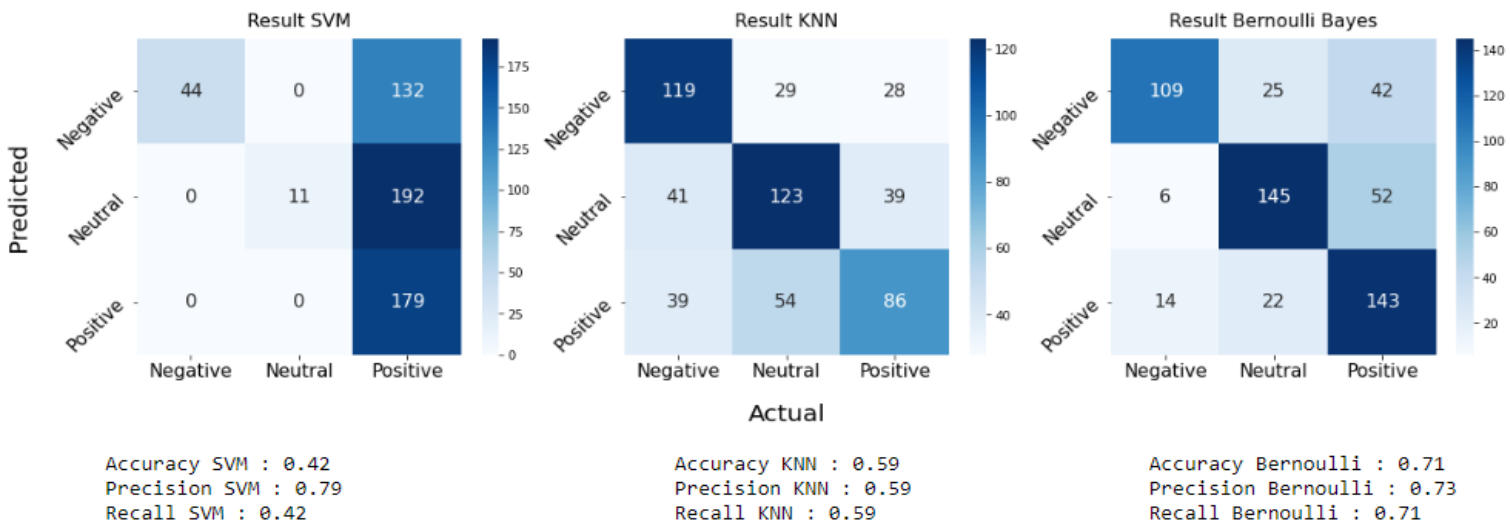
1. Pengecekan data kosong (NaN) => Menghasilkan 280 data bersatus NaN
2. Split value pada kolom Time => Menghasilkan dua kolom baru At_Time dan Date
3. Mengambil dua kolom => Kolom 'Headlines' dan 'Description' untuk dilatih
Note : Pemilihan diambil dua kolom tersebut untuk project sentiment analisis saat ini
4. Pengecekan data duplicated => Terdapat 10 data duplicated
5. Penghapusan data NaN dan duplicated => Data bersih berjumlah 2790 dari 3080.
6. Text preprocessing => case folding, tokenizer, stopwords, stemmer, dan lemmer
7. Decision label menggunakan SentimentIntensityAnalyzer => Terdapat 3 label (Positive, Negative, Neutral)
8. Modeling menggunakan pipeline => train_test_split, Count vector, TF-IDF, modeling (SVM (kernel RBF), KNN (N = 5 (Automatic)), Bernoulli Naïve Bayes)

Saran

1. Dapat dilakukan TF hanya untuk Bernoulli Naïve Bayes dikarenakan bersifat probability dan melakukan balancing label (teknik diantaranya yaitu SMOTE) untuk menghasilkan performa yang lebih baik pada setiap model
2. SVM dapat mencoba kernel linear, dikarenakan kernel linear pada SVM merupakan kernel yang memungkinkan untuk permasalahan text classifier
3. Menggunakan Cross Validation untuk pembagian data train dan data testing

AFTER PREPROCESSING

	Headlines	Description	label
0	jim cramer better way invest covid vaccin gold...	mad money host jim cramer recommend buy four c...	Positive
1	cramer lightn round would teradyn	mad money host jim cramer ring lightn round be...	Negative
2	cramer week ahead big week earn even bigger we...	well pay earn noncovid compani lancet publish ...	Positive
3	iq capit ceo keith bliss say tech healthcar ralli	keith bliss iq capit ceo join close bell talk ...	Positive
4	wall street deliv kind pullback ive wait jim c...	look stock highqual compani go lower even thou...	Negative
...
2785	market lack christma cheer	accord kensho here market fare week christma p...	Neutral
2786	cramer remix biggest mistak make tax stock mar...	jim cramer reveal top rule come tax sell stock	Positive
2787	cramer say own mani stock littl cash set failur	jim cramer broke own fewer stock help make mon...	Negative
2788	cramer help investor flash crash follow one ke...	jim cramer built nobodi ever made dime panick ...	Neutral
2789	cramer never buy stock youll almost definit ge...	jim cramer doubl key invest rule never buy des...	Neutral



Kesimpulan

1. Model Bernoulli Naïve Bayes menghasilkan akurasi yang cukup baik dan dari nilai presisi serta recall tidak searah dengan nilai akurasi yang dihasilkan
2. SVM dan KNN memiliki akurasi < 70. Hal ini dapat terjadi dikarenakan ketidak seimbangan data yang dimili, dikarenakan SVM dan KNN mengaldakan hyperlane dan neighbour untuk menentukan classifier.
3. Kategori positif menjadi kategori yang paling tinggi dalam tingkat ketepatan prediksi, hal ini dapat diperkirakan dikarenakan label positif bersifat 'mayoritas'