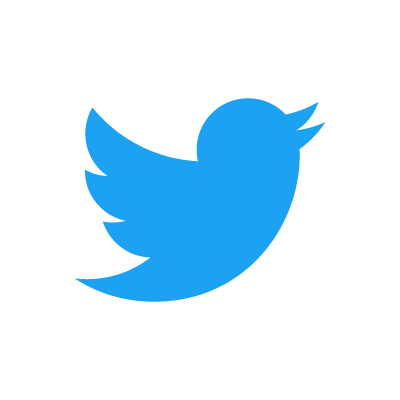
Sentiment Analysis of Tweets



Oleh :

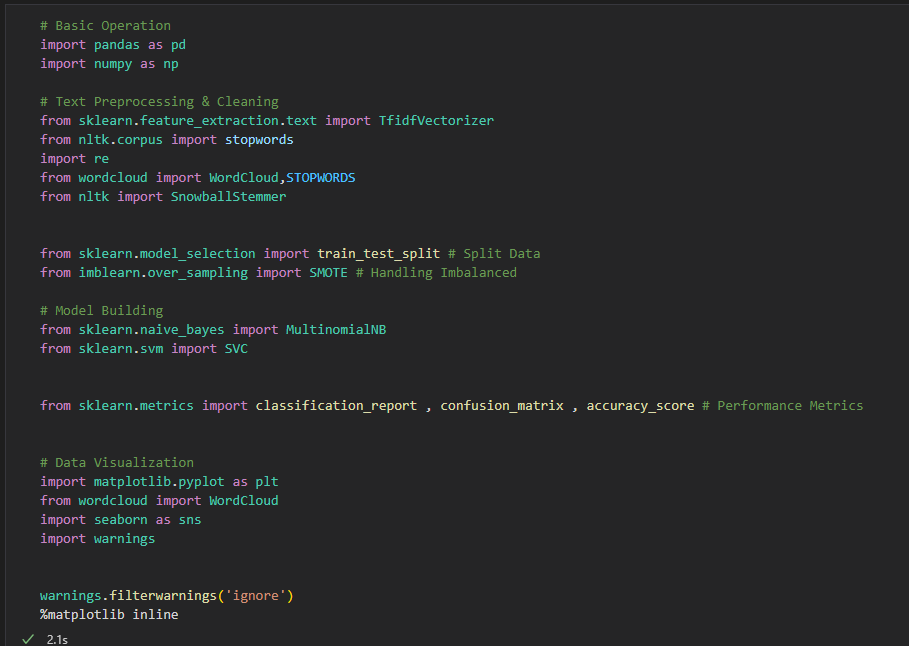
Dhoni Hanif Supriyadi

1. **Business Understanding**

Pada tahapan ini, seorang Data Scientist akan menelaah bisnisnya terlebih dahulu seperti apa yang dibicarakan dalam bisnis ini? Apa bisnis ini termasuk klasifikasi atau regresi? Dan hal lainnya. Bisnis ini membicarakan tentang Sentiment Analysis terhadap twitter. Seperti yang kita ketahui, beberapa kata tidak dapat ditolerir seperti kata – kata yang negative sehingga hal ini butuh kita nilai bahwa kata – kata ini buruk. Sebelum itu, kita harus mengumpulkan datanya terlebih dahulu dan menganalisisnya terlebih dahulu.

1. **Data Collection**

Hal pertama yang harus kita lakukan adalah mengumpulkan datanya terlebih dahulu. Disini, kita sudah mendapatkan datanya melalui kaggle.com. Lalu, kita akan lakukan pembacaan datanya. Namun, sebelum itu, mari import library yang dibutuhkan terlebih dahulu.

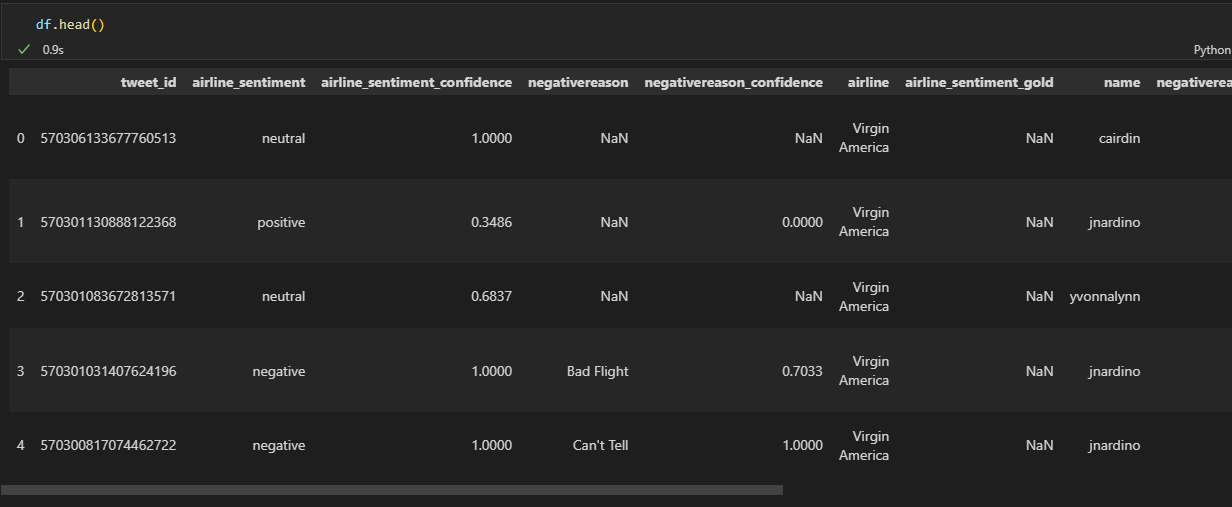


Disini, kita membutuhkan beberapa library seperti pandas, numpy, scikit-learn, nltk, re, wordcloud, seaborn, warnings, dan matplotlib.

Selanjutnya, mari lakukan pembacaan data seperti berikut.

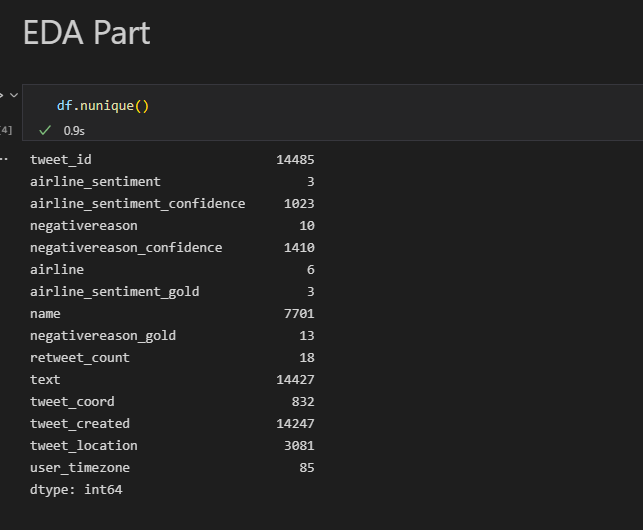


Data berhasil diload. Sekarang, cek 5 data dari atas.

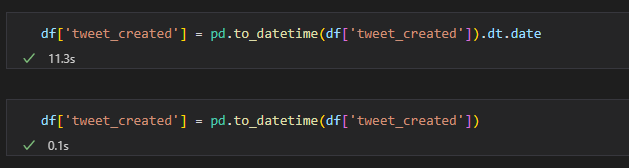


1. **Exploratory Data Analysis (EDA)**

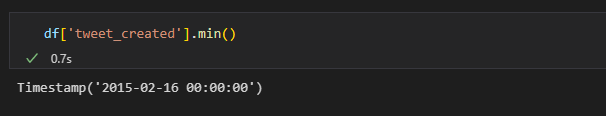
Hal pertama yang harus kita lakukan adalah mengecek jumlah nilai unique seperti berikut.



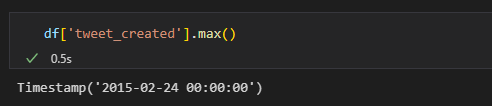
Disini dikatakan bahwa tweet\_id, text, dan tweet\_created memiliki nilai unik yang sangat besar. Seperti yang kita ketahui, ada beberapa variabel yang menyimpan data dengan format yang salah seperti variabel tweet\_created. Oleh karena itu, tweet\_created memiliki jumlah nilai unik yang sangat besar sehingga kita harus memperbaiki format data tersebut seperti berikut.



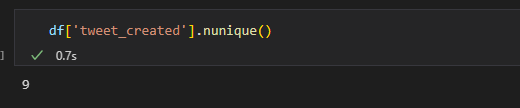
Data minimal dari tweet\_created adalah sebagai berikut.



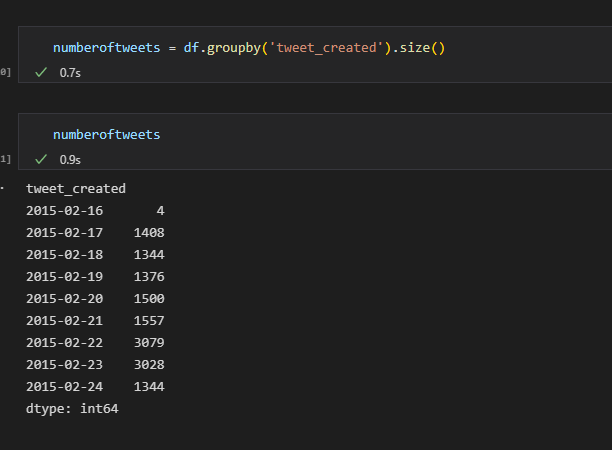
Sedangkan data maksimal dari tweet\_created adalah sebagai berikut.



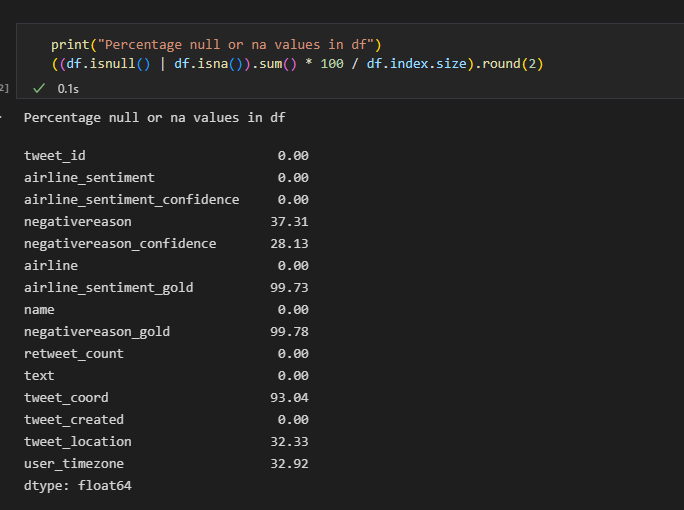
Lalu, mari cek kembali jumlah nilai unik dari variabel tweet\_created seperti berikut.



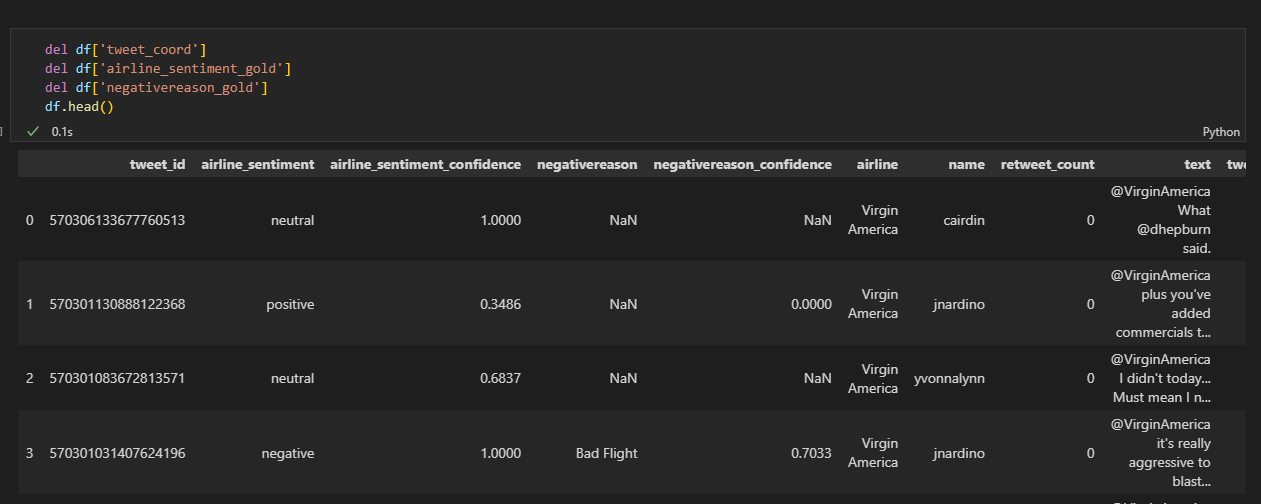
Setelah kita perbaiki format data, maka jumlah nilai unik berubah menjadi 9. Selanjutnya, mari cek jumlah tweets berdasarkan tweet\_created seperti berikut.



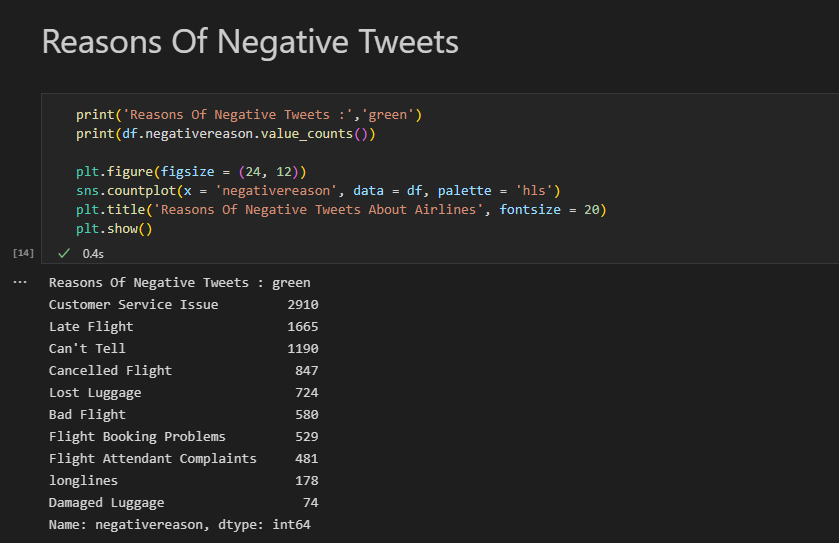
Data terbesar ada pada tanggal 22 February 2015, sedangkan data terkecil berada pada tanggal 16 February 2015. Sekarang, mari cek missing values seperti berikut.

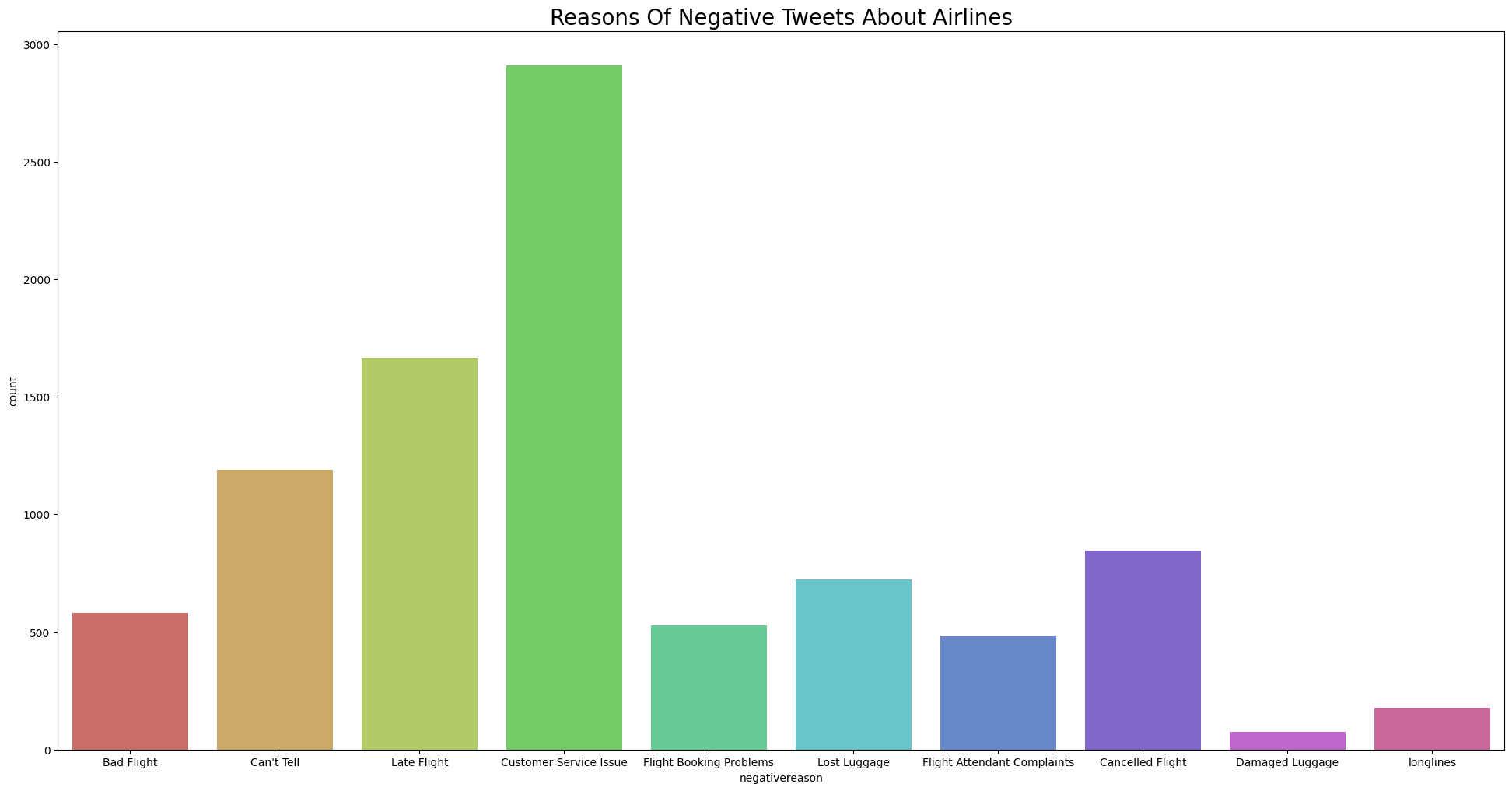


Dalam hal ini, kita menampilkan missing value dari masing – masing data dalam bentuk persentase. Missing value terbesar berada pada data negativereason\_gold yaitu sebesar 99.78%. beberapa data juga memiliki missing value yang sangat besar seperti airline\_sentiment\_gold dan tweet\_coord. Mari hapus data – data ini.

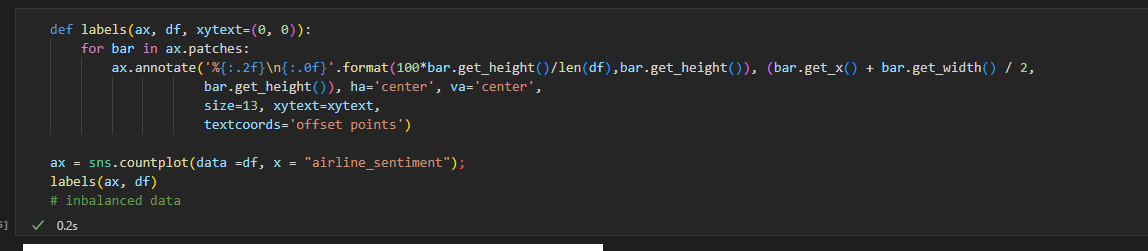


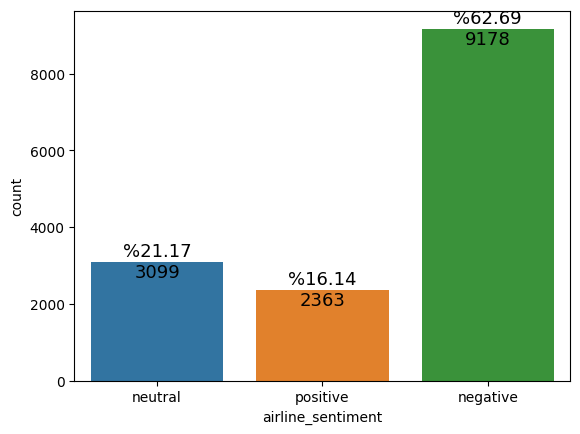
Sekarang, data – data yang memiliki missing value yang sangat banyak telah dihapus. Selanjutnya, mari tampilkan jumlah dari nilai negative reason seperti berikut.





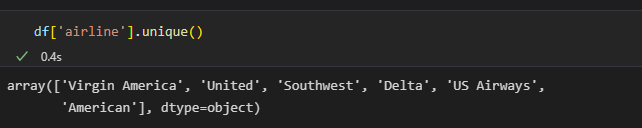
Seperti yang kita lihat, data terbesar berada pada Customer Service Issue yaitu sebesar 2910 data dan data terkecil berada pada Damage Luggage yaitu sebesar 74 data. Selanjutnya, mari tampilkan label dari data.



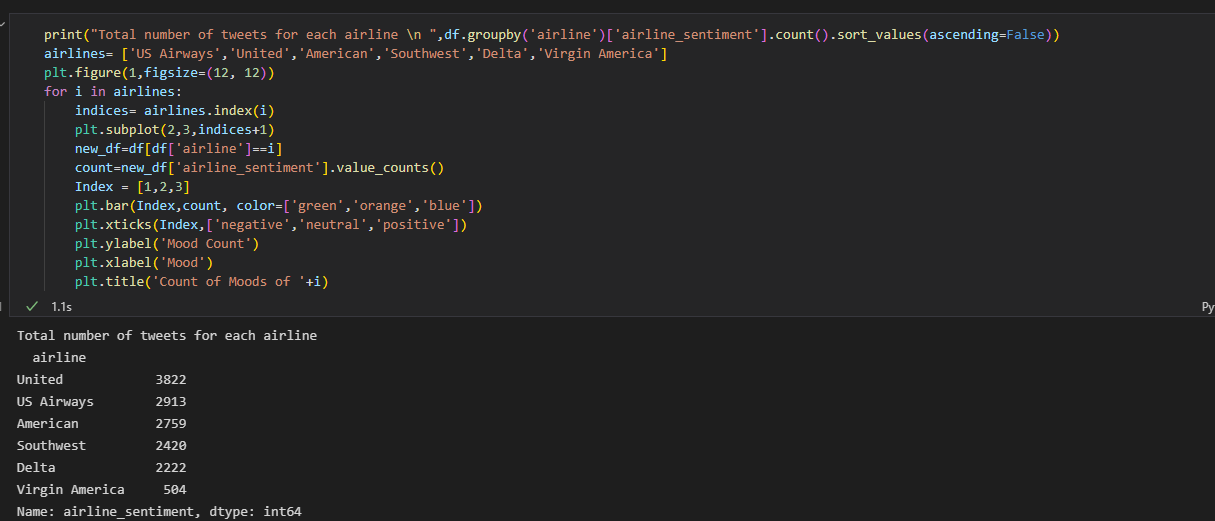


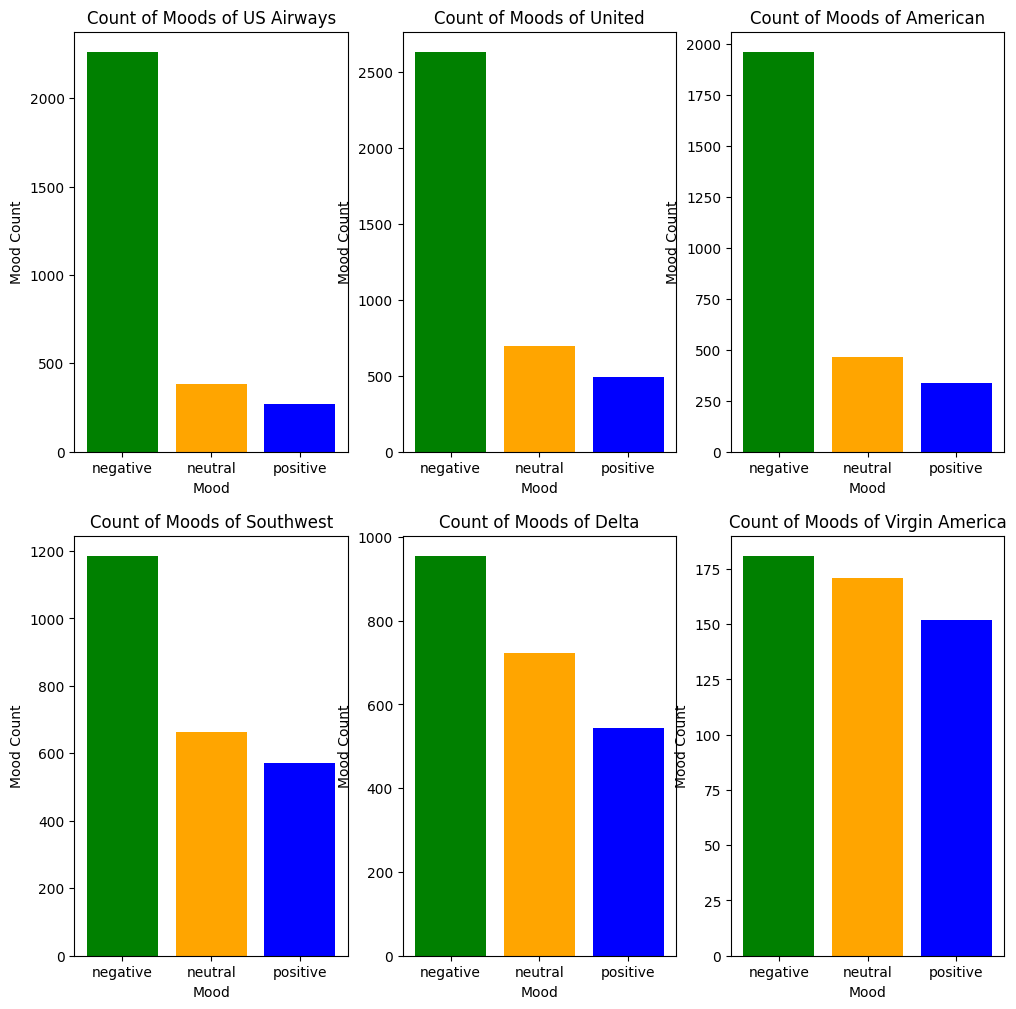
Seperti yang terlihat, data negative memiliki angkat yang sangat besar yaitu 62.69 % dari keseluruhan data sedangkan data positif hanya 16.14 % dan data neutral hanya 21.17 % sehingga dapat kita katakana bahwa data inblance.

Selanjutnya, mari tampilkan nilai unik dari data airline.

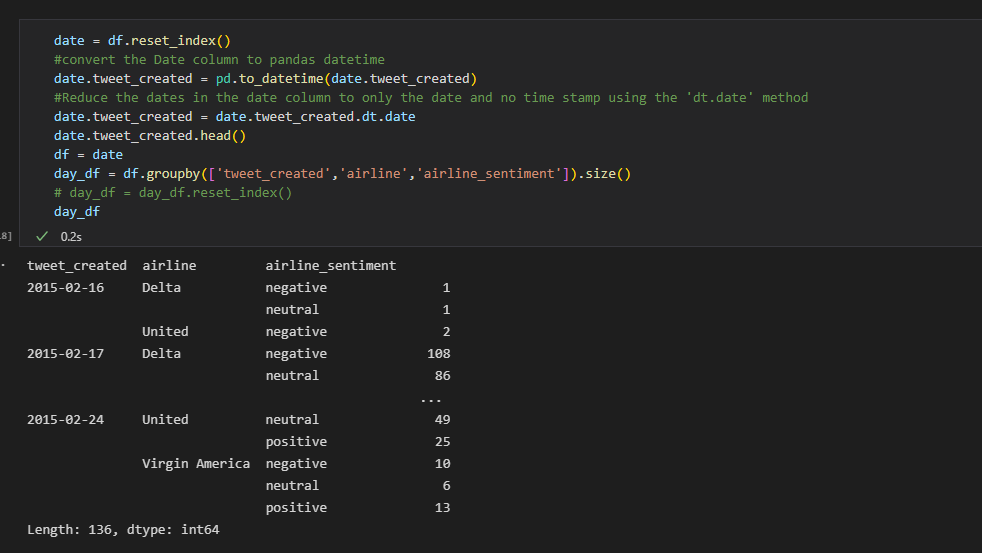


Data airline memiliki nilai unik yaitu seperti yang terlihat diatas. Lalu, tampilkan data tersebut berdasarkan mood seperti berikut.

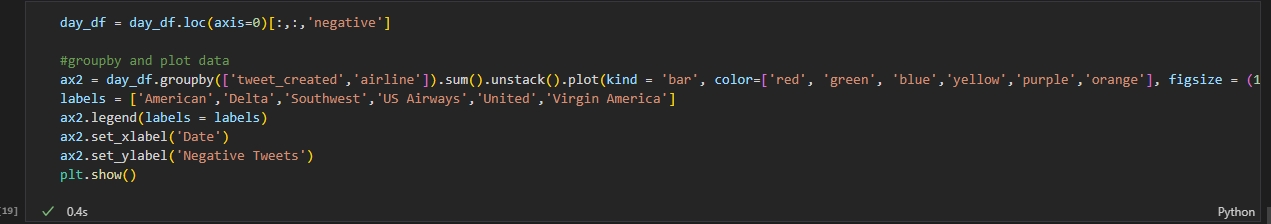


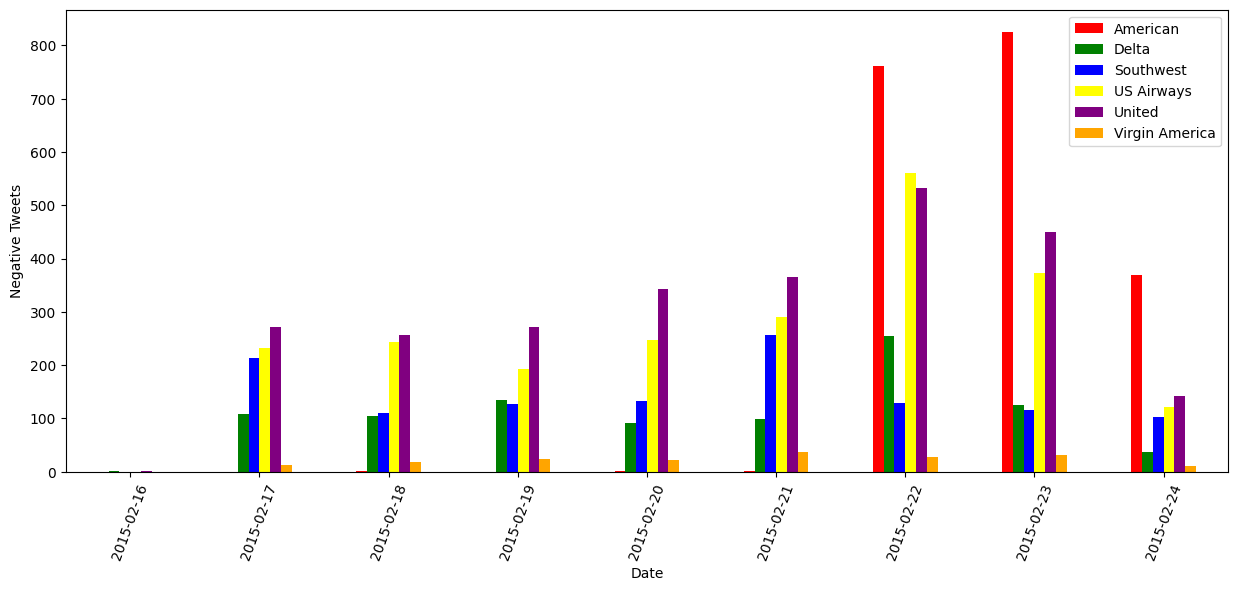


Seperti yang terlihat, data negative adalah data terbesar di semua tempat yang terdaftar sedangkan data positif memiliki angka yang paling rendah. Data terbesar juga dimiliki oleh United sedangkan data terkecil dimiliki oleh Virgin America. Selanjutnya mari ubah index menjadi tweet\_created.



Kita telah berhasil mengubah index menjadi tanggal. Selanjutnya, mari tampilkan data tersebut berdasarkan tanggal dan airline.

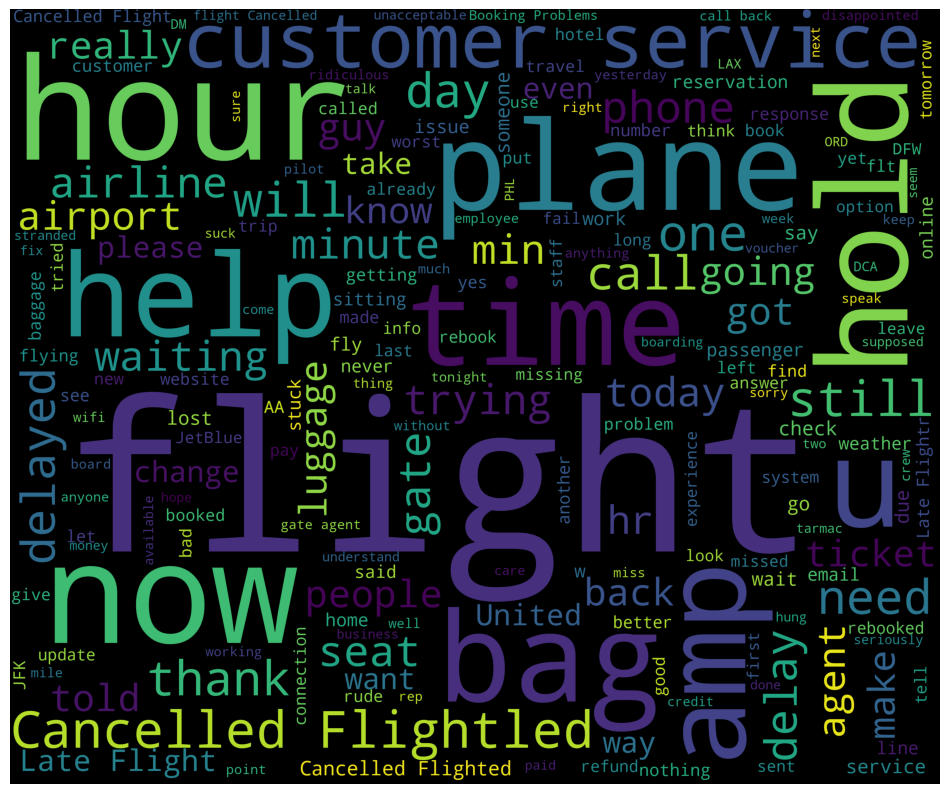




Seperti yang kita lihat, data pada tanggal 23 Februari 2015 memiliki data paling besar disbanding seluruh data. Amerika menduduki data paling tinggi pada tanggal 22, 23, dan 24 Februari 2015. Namun, pada tanggal lain, amerika menduduki angka paling rendah. United memiliki kencederungan sama dari tanggal ke tanggal tetapi mengalami penurunan pada tanggal 23 dan 24 Februari 2015. Data virgin adalah data yang paling rendah dari hampir seluruh data di beberapa tanggal.

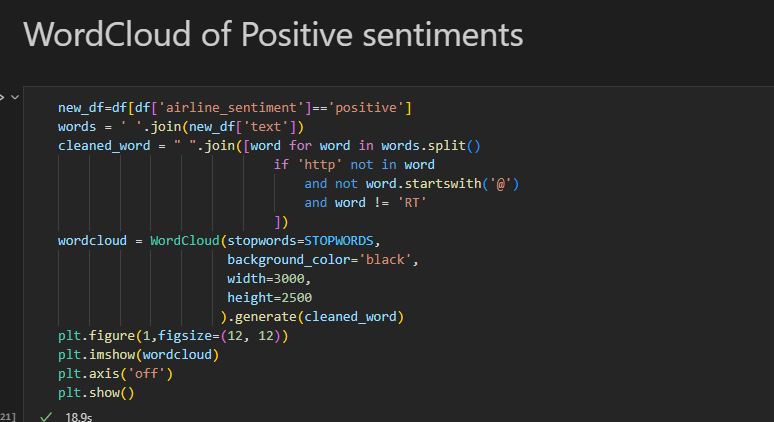
Selanjutnya, mari tampilkan wordcloud pada data negative seperti berikut.

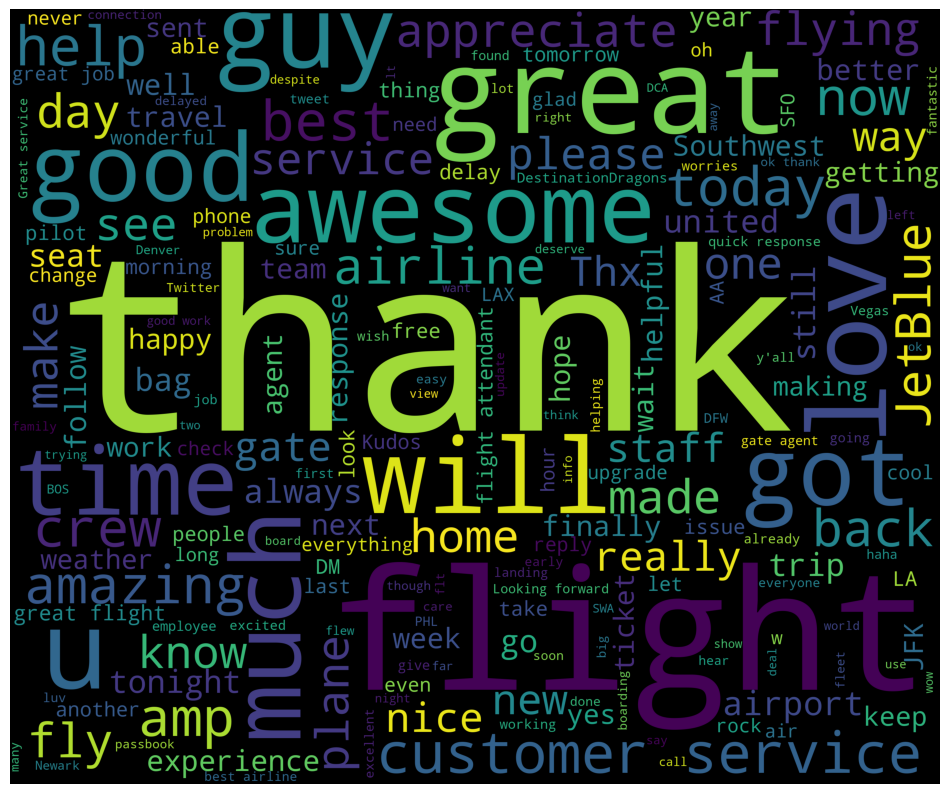




Seperti yang kita lihat, semakin banyak kata yang sama pada data, semakin besar pula kata tersebut. Disini, terlihat kata flight adalah kata yang sangat besar sehingga dapat kita katakan bahwa kata flight mewakili hampir seluruh data. Namun, beberapa kata terlihat sangat kecil yang artinya hanya ada sedikit data mengenai hal tersebut pada negative wordcloud.

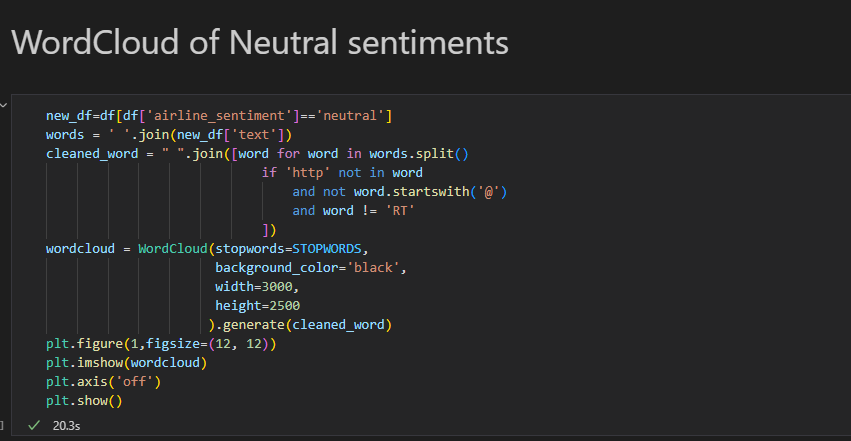
Selanjutnya, mari tampilkan wordcloud pada positive sentiments

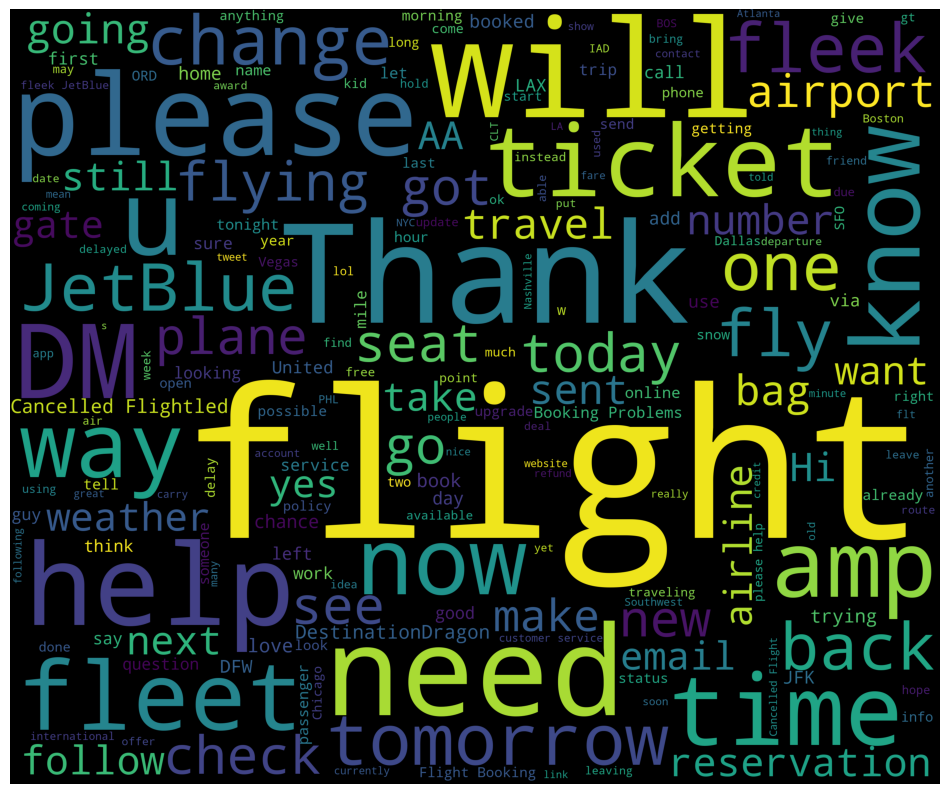




Seperti yang terlihat, kata thank adalah kata yang paling besar sehingga dapat kita katakan bahwa kata thank mewakili hampir seluruh data positive. Namun, kata flight juga termasuk kata yang besar. Sama halnya seperti pada data negative, kata flight pada positive juga sangat banyak.

Selanjutnya, mari tampilkan wordcloud pada neutral sentiments.

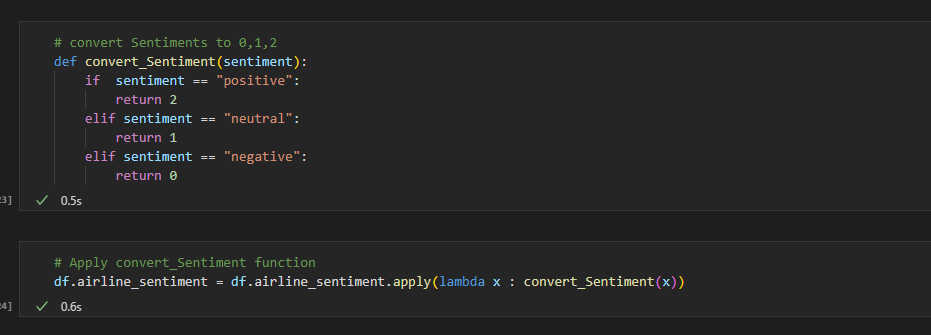




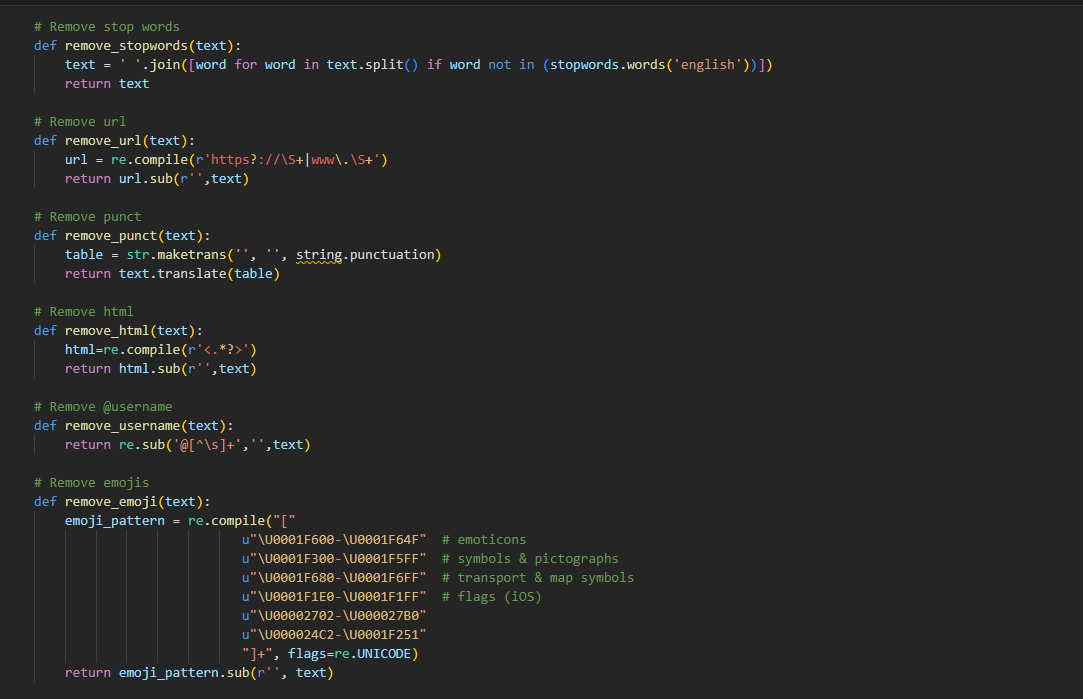
Seperti yang kita lihat, kata flight disini juga mewakili hampir seluruh data neutral sehingga dapat kita simpulkan bahwa kata flight mewakili hampir seluruh data pada data negative, positive, dan neutral.

1. **Data Preparation**

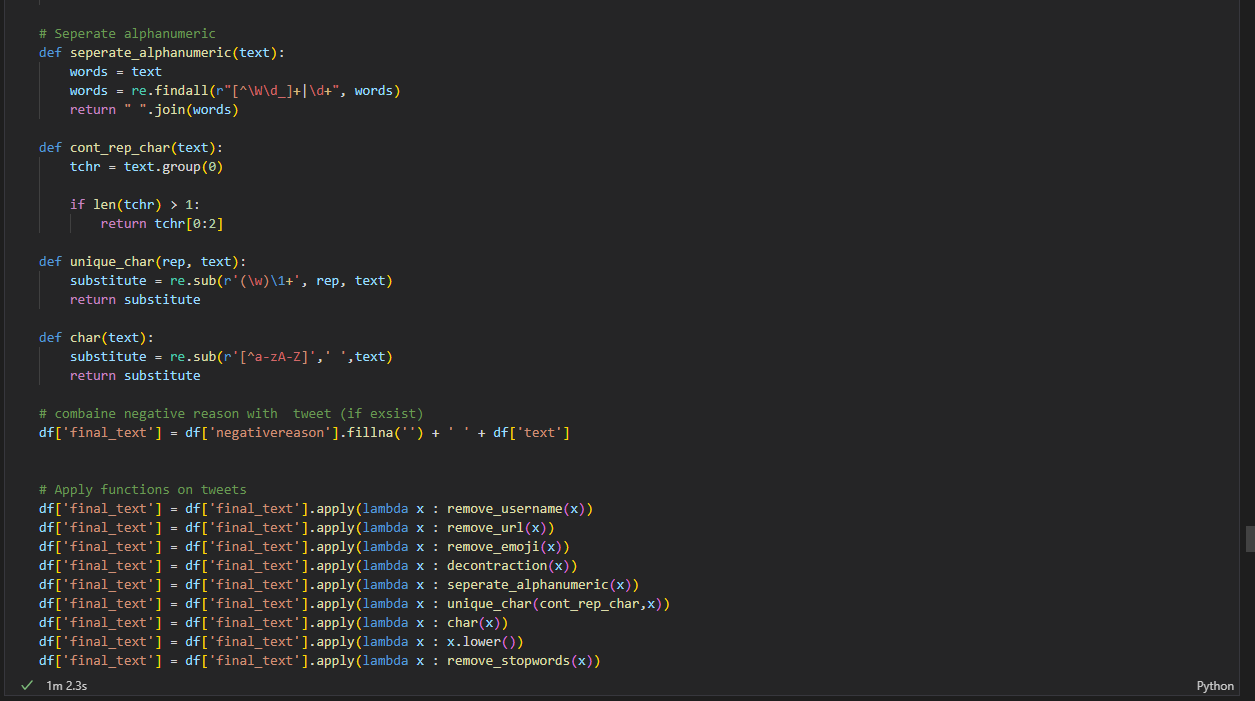
Pada tahapan ini, kita akan menyiapkan datanya terlebih dahulu dan melakukan beberapa tahapan terhadap data seperti stopwords, filtering, stemming, dan vektorisasi sebelum melatihnya pada machine learning. Pertama, kita ubah data negative, positive dan neutral menjadi bilangan diskrit seperti berikut.



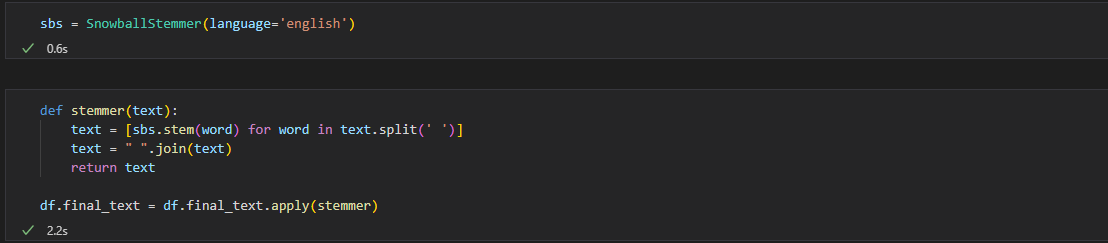
Lalu, lakukan stopwords pada data seperti berikut.



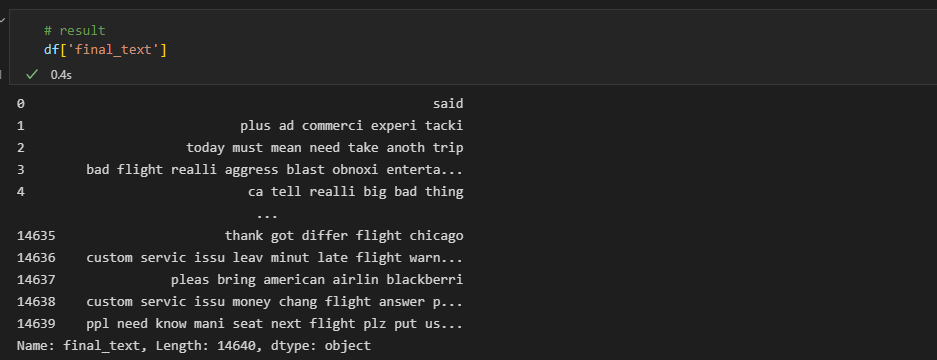




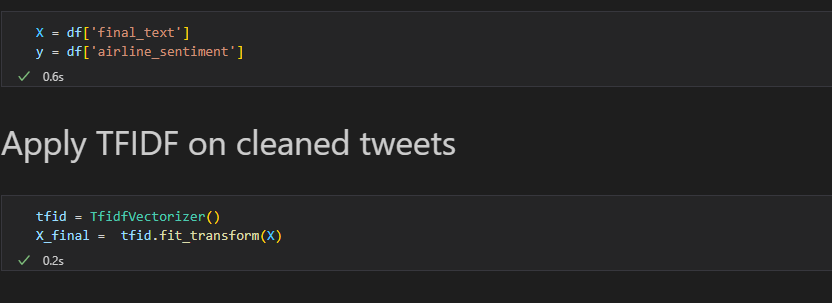
Kita berhasil melakukan stopwords pada data. Selanjutnya, melakukan stemming seperti berikut.



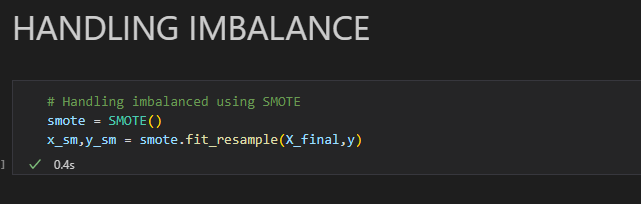
Dan mengecek hasilnya seperti berikut.



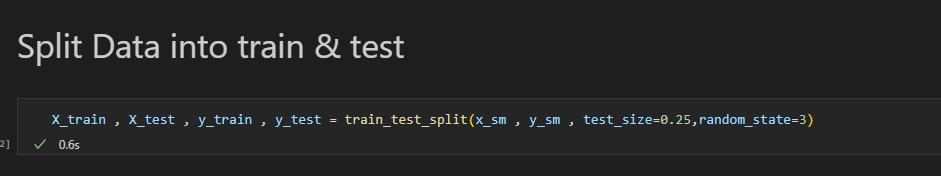
Data berhasil kita stemming. Selanjutnya, kita lakukan vektorisasi pada data dengan TfidfVectorizer().



Data berhasil divektorisasi. Selanjutnya, lakukan smote pada data guna meng-handle data imbalance atau data tidak imbang seperti berikut.



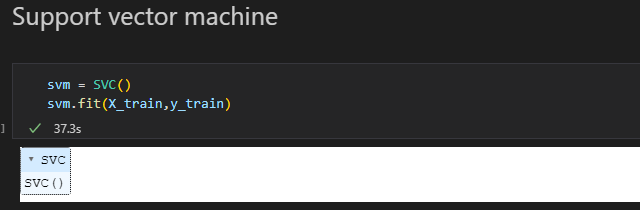
Lalu, bagi dataset menjadi 2 yaitu data training dan data testing seperti berikut.



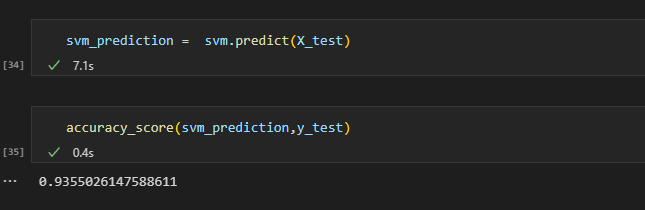
Data telah dibagi menjadi data training dan data testing dengan rasio 75:25 dan random\_state 3. Selanjutnya, kita lakukan tahap pembuatan model.

1. **Building Model & Evaluation Model**

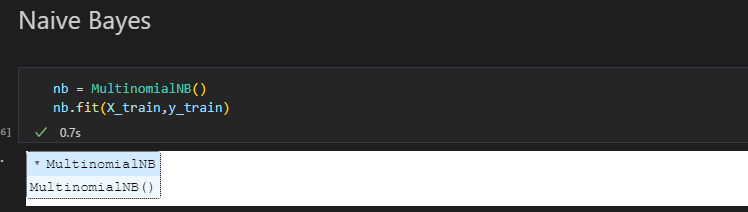
Pada tahapan ini, kita lakukan pembuatan model dan mengevaluasinya. Kita memakai 2 model yaitu Support Vector Machine (SVM) dan Naïve Bayes. Pertama, kita buat model Support Vector Machine (SVM) seperti berikut.



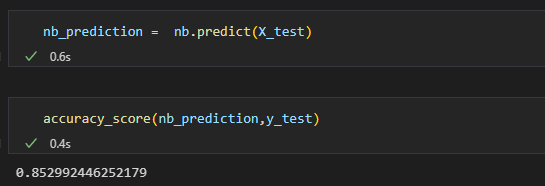
Lalu, cek akurasi pada model ini seperti berikut.



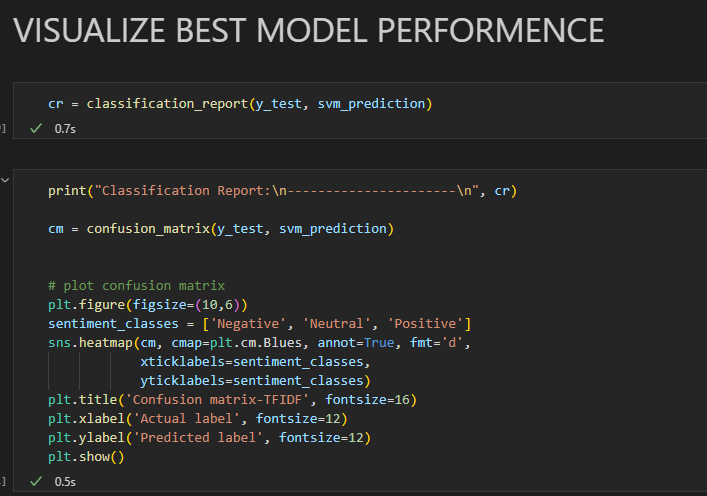
Model menghasilkan akurasi 93 % pada data testing. Selanjutnya lakukan pembuatan model Naïve Bayes seperti berikut.

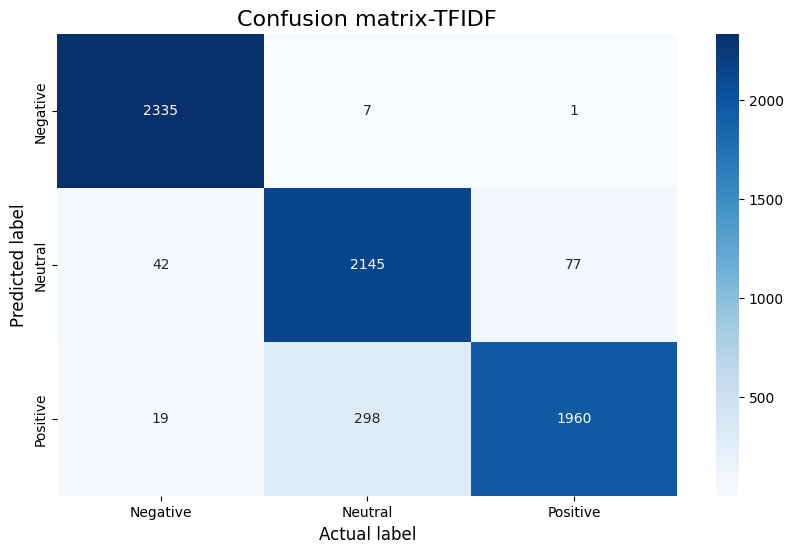
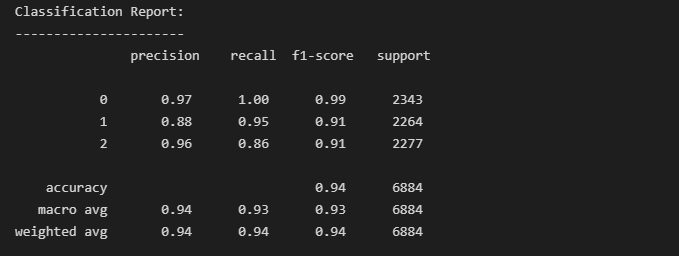


Lalu, cek akurasi pada model ini.



Model ini memiliki akurasi yang lebih rendah dibanding model Support Vector Machine (SVM) sehingga kita dapat menggunakan model Support Vector Machine (SVM). Sebelum itu, mari evaluasi model Support Vector Machine (SVM) seperti berikut.





Seperti yang kita lihat, model memiliki akurasi sekitar 94% dengan True Positive jauh lebih besar dibanding True Negative dan begitupun juga True Negative sehingga kita dapat memakai model ini untuk kebutuhan selanjutnya.

Terimakasih