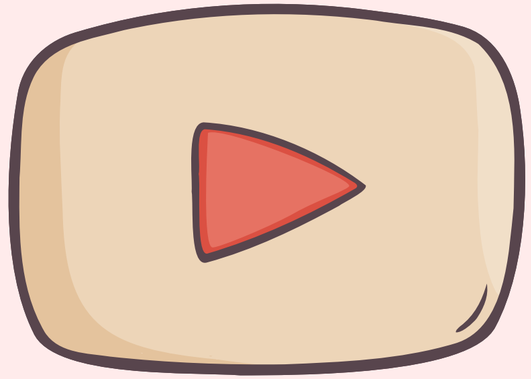
SPAM DETECTION ON YOUTOBE COMMENTS

PRESENT BY:RIFQI ARRAYAN



BACKGROUND

The YouTube platform has become one of the most popular mediums for digital content consumption. Many companies are utilising YouTube as an effective marketing tool to reach a wide audience. One aspect that is important for companies looking to advertise is maintaining their brand image and the effectiveness of their marketing campaigns. Spam comments under videos are a significant problem as they can ruin the user experience and reduce the value of the content, which in turn can affect the perception of the advertised brand.

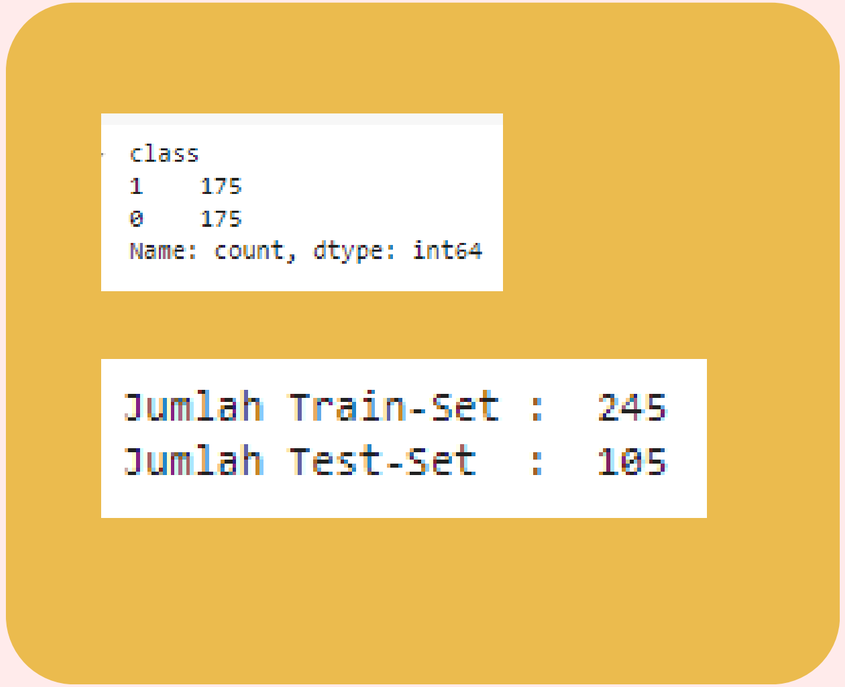
DATA LOADING

	comment_id	author	date	content	class
0	LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: ...	True
1	LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A	adam riyati	2013-11-07T12:37:15	Hey guys check out my new channel and our firs...	True
2	LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	True
3	z13jhp0bxqncu512g22wvzkasxmvvzjaz04	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^_^	True
4	z13fwbwp1oujthgqj04chlngpvzmtt3r3dw	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtWQ Check this out .	True
5	LZQPQhLyRh9-wNRtlZDM90f1k0BrdVdJyN_YsaSwfxc	Jason Haddad	2013-11-26T02:55:11	Hey, check out my new website!! This site is a...	True
6	z13lfzdo5vmdi1cm123te5uz2mqig1brz04	ferleck ferles	2013-11-27T21:39:24	Subscribe to my channel	True
7	z122wfnzgt30fhubn04cdn3xfx2mxzngsl40k	Bob Kanowski	2013-11-28T12:33:27	i turned it on mute as soon is i came on i jus...	False
8	z13ttt1jcraqexk2o234ghbgzxymz1zzi04	Cony	2013-11-28T16:01:47	You should check my channel for Funny VIDEOS!!	True
9	z12avveb4xqiirsix04chxviiljryduwxg0	BeBe Burkey	2013-11-28T16:30:13	and u should.d check my channel and tell me wh...	True

	comment_id	author	date	content	class
0	LZQPQhLyRh80UYxNuaDWhIGQYNQ96IuCg-AYWqNPjpU	Julius NM	2013-11-07T06:20:48	Huh, anyway check out this you[tube] channel: ...	1
1	LZQPQhLyRh_C2cTtd9MvFRJedxydaVW-2sNg5Diuo4A	adam riyati	2013-11-07T12:37:15	Hey guys check out my new channel and our firs...	1
2	LZQPQhLyRh9MSZYnf8djyk0gEF9BHDPYrrK-qCczIY8	Evgeny Murashkin	2013-11-08T17:34:21	just for test I have to say murdev.com	1
3	z13jhp0bxqncu512g22wvzkasxmvvzjaz04	ElNino Melendez	2013-11-09T08:28:43	me shaking my sexy ass on my channel enjoy ^_^	1
4	z13fwbwp1oujthgqj04chlngpvzmtt3r3dw	GsMega	2013-11-10T16:05:38	watch?v=vtaRGgvGtWQ Check this out .	1
5	LZQPQhLyRh9-wNRtlZDM90f1k0BrdVdJyN_YsaSwfxc	Jason Haddad	2013-11-26T02:55:11	Hey, check out my new website!! This site is a...	1
6	z13lfzdo5vmdi1cm123te5uz2mqig1brz04	ferleck ferles	2013-11-27T21:39:24	Subscribe to my channel	1
7	z122wfnzgt30fhubn04cdn3xfx2mxzngsl40k	Bob Kanowski	2013-11-28T12:33:27	i turned it on mute as soon is i came on i jus...	0
8	z13ttt1jcraqexk2o234ghbgzxymz1zzi04	Cony	2013-11-28T16:01:47	You should check my channel for Funny VIDEOS!!	1
9	z12avveb4xqiirsix04chxviiljryduwxg0	BeBe Burkey	2013-11-28T16:30:13	and u should.d check my channel and tell me wh...	1

in the picture on the side, I transformed the label (class) to change it from string data to numeric form so that it can be read by machine learning.

DATA LOADING



in the picture on the side I checked the number of labels or classes, where the class with label 1 (spam) is 175 rows and the class with label 0 (not spam) is 175 rows.
as well as displaying the amount of data included in the training set (train set) 70% and the test set (test set) 30%.

VEKTORISASI

```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer()
```

```
# Mempelajari kata apa saja yang ada pada `train_comment`
vect.fit(train_comment)

# Mengubah kata menjadi vektor berdasarkan frekuensi kemunculannya
train_comment_mat = vect.transform(train_comment)
test_comment_mat = vect.transform(test_comment)
```

Convert words to vectors based on their frequency of occurrence, and use the transfrom method to convert test data and test data to numeric based on their occurrence.

KLASIFIKASI

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5)
```

```
knn.fit(train_comment_mat, train_class)
```

```
from sklearn.metrics import accuracy_score

hasil_prediksi = knn.predict(test_comment_mat)
print('Akurasi : ', accuracy_score(test_class, hasil_prediksi))
```

```
Akurasi :  0.8952380952380953
```

performs classification with the K-Nearest Neighbors (KNN) technique by considering 5 nearest neighbours to determine the class of a data point and checking between The first parameter is the actual label, and the second parameter is the predicted label.with an accuracy rate close to 90%.

NEW COMMENT TEST

```
komentar_baru = ["saya Rifqi arrayan sangat ganteng",
                 "Subscribe to me for daily vines"]

# Mengubah teks ke dalam bentuk vector
komentar_baru = vect.transform(komentar_baru)

# Memprediksi komentar baru
hasil_prediksi_komentar_baru = knn.predict(komentar_baru)
print('Hasil Prediksi : ', hasil_prediksi_komentar_baru)
```

```
Hasil Prediksi :  [0 1]
```

in the picture beside I did a test on a new comment, where I tried a test with me ‘Rifqi arrayan sangat ganteng’ and the machine guessed that this comment was not spam (0) then I also tested with ‘Subscribe to me for daily vines’ and the machine guessed that this comment was spam (1)

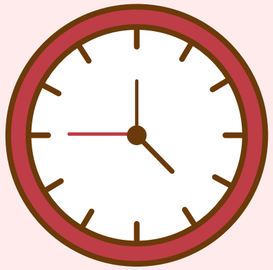
CONCLUSION

Spam detection analysis of YouTube comments is an important step for companies looking to advertise on the platform, as spam comments can ruin user experience and create a negative image for brands. Using machine learning techniques and natural language processing, companies can effectively identify and filter spam, maintaining the quality of interactions on the channels chosen for advertising. This ensures that the company's ads appear in a positive and relevant context, increasing audience trust and engagement, and optimising the effectiveness of marketing campaigns.

ATTENTION
PLEASE!

VIEW IN COLAB

VIEW IN GITHUB



THANKS YOU