

# Lab 02 Supervised Learning

## 'Hate Speech Classifier'

Bogumila Dubel | Johanthä Hänni | Yannic Wetzel

### Einleitung:

In dem Lab wurden drei verschiedene Modelle auf den zur Verfügung gestellten Testdaten trainiert. Im ersten Schritt wurden die Scores der verschiedenen Ansätze miteinander verglichen. Anschliessend wurde jedes der Modelle auf einem Set von 'Review' - Daten, die dem Modell unbekannt sind, auf einer 'hate\_speech\_predicat' - Applikation ausgeführt. Zur Optimierung der Laufzeit der 'hate\_speech\_predicat' - Applikationen wurden die trainierte Modelle jeweils gespeichert und in den Applikationen zur Wiederverwendung geladen.

Aus den unten geschilderten Ergebnissen geht es klar hervor, dass der Deep Learning Ansatz mit dem Keras Modell die besten Ergebnisse liefert.

### Linear Support Vector Classification (Baseline):

#### Einführung:

Es handelt sich um ein Basismodell, welches im Lab schon enthalten war und auf dem Linear Support Vector Classification basiert. Das Modell ist in der Lage die gegebenen Daten in zwei Klassen 0 = hate comment und 1 = no hate comment aufzuteilen.

#### Tool:

- LinearSVC von scikit-learn

#### Scores / Beobachtungen:

Classification and evaluation					
	precision	recall	f1-score	support	
0	0.97	0.90	0.94	5734	
1	0.48	0.79	0.59	649	
accuracy			0.89	6383	
macro avg	0.73	0.84	0.77	6383	
weighted avg	0.92	0.89	0.90	6383	
[[5177 557]					
[ 139 510]]					
Using 3 folds					

Das Modell liefert rechte gute Ergebnisse mit einer Test accuracy von 89%, es gibt aber eindeutig Verbesserungspotential. Im Anhang 'baseline' kann man schon in dem kleinen Set ein paar Einträge finden, die eindeutig kein 'hate comment' sind, wurden jedoch als solches eingestuft wie z.B. 'what do you want to know?'

#### Resultate:

Text-Document: 'hate\_speech\_predict\_baseline.txt' oder Anhang 'baseline'

## Logistic Regression:

### Einführung:

Ähnlich wie LinearSVC kann man mithilfe von Logistic Regression die Daten nur in zwei Klassen (0 = hate comment und 1 = no) aufteilen. Anhand von Resultaten kann man feststellen, dass es jedoch ein geeigneteres Modell für diesen Aufgabentyp ist.

### Tool:

- LogisticRegression von scikit-learn

### Scores / Beobachtungen:

	precision	recall	f1-score	support
0	0.95	1.00	0.97	5734
1	0.96	0.53	0.68	649
accuracy			0.95	6383
macro avg	0.95	0.77	0.83	6383
weighted avg	0.95	0.95	0.94	6383
[[5718 16]				
[ 303 346]]				

Das Modell liefert deutlich bessere Ergebnisse als LinearSVC mit einer Test accuracy von 95%. Im Anhang 'regression' werden in dem kleinen Set keine Einträge gefunden, die auffällig unpassend eingeteilt wurden.

### Resultate:

Text-Dokument: 'hate\_speech\_predict\_regression.txt' oder Anhang regression

## Deep Learning CNN Keras:

### Einführung:

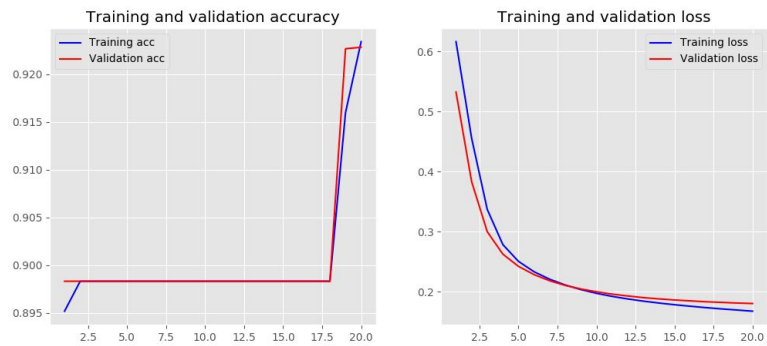
Im Gegensatz zu den Machine Learning Modellen kann man die Daten in mehrere Klassen als nur in zwei aufteilen. Vorteilhaft ist auch die Tatsache, dass die Wahrscheinlichkeit der Zugehörigkeit zu jeder Klasse mitgeliefert wird. Auf diese Art und Weise kann man diese Kommentare herausfiltern, die zu keiner Klasse eindeutig zugewiesen werden können.

### Vorgehen:

Im ersten Schritt wurde nach einem passenden Modell gesucht. Die folgenden Einstellungen wurden dabei justiert: Dense der Layers, Anzahl der Iterationen, Batch Size. Das Modell wurde durchs Plotten und Analyse der 'learning curve' ausgewählt. Alle Modelle basieren auf 2 Layers. Hinzufügen zusätzlicher Layers hat keine bedeutende Verbesserung gebracht und das Trainieren wesentlich verlangsamt. Beiliegend die Resultate.

Dense: 4  
Iterationen: 20  
Batch Size: 512

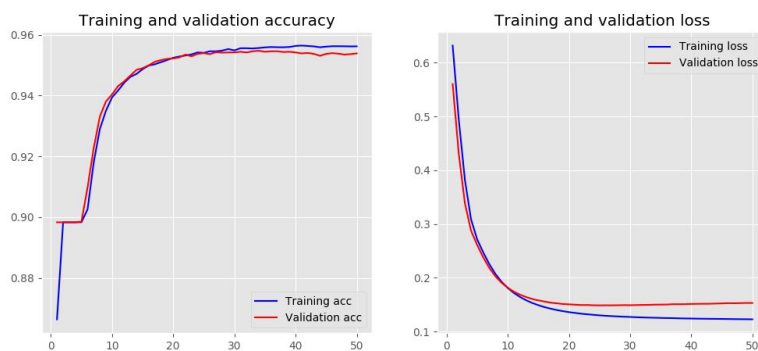
Training Accuracy:  
0.9232  
Validation Accuracy:  
0.9220



Die Ergebnisse sind schlechter als die, welche vom Logistic Regression geliefert wurden.

Dense: 4  
Iterationen: 50  
Batch Size: 512

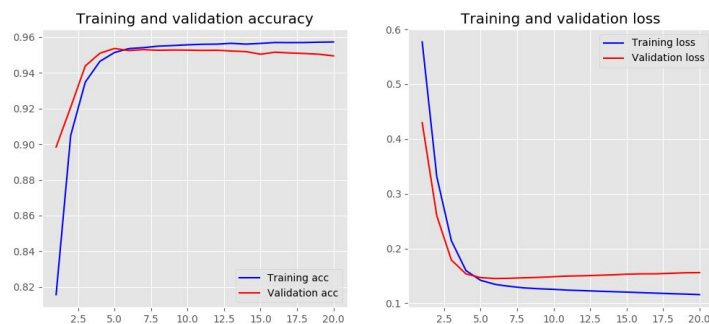
Training Accuracy:  
0.9565  
Validation Accuracy:  
0.9539



Das Erhöhen der Anzahl der Iterationen steigert die Accuracy um ca. 3%.

Dense: 16  
Iterationen: 20  
Batch Size: 256

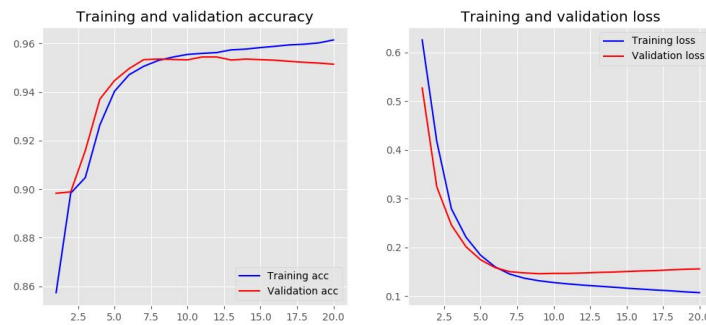
Training Accuracy:  
0.9591  
Validation Accuracy:  
0.9496



Keine deutliche Verbesserung gegenüber dem oberen Modell. Zusätzlich gehen die Kurven der Training and Validation Accuracy leicht auseinander.

Dense: 16  
Iterationen: 20  
Batch Size: 512

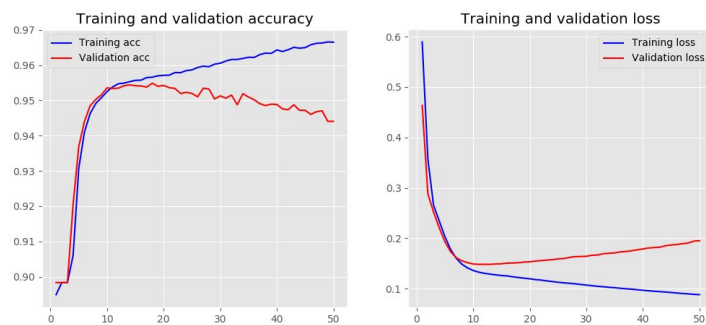
Training Accuracy:  
0.9632  
Validation Accuracy:  
0.9514



Eine leichte Verbesserung ist zu beobachten. Die Tendenz sich voneinander entfernenden Kurven kann beobachtet werden, welche jedoch noch nicht als 'Overfitting' eingestuft werden kann.

Dense: 16  
Iterationen: 50  
Batch Size: 512

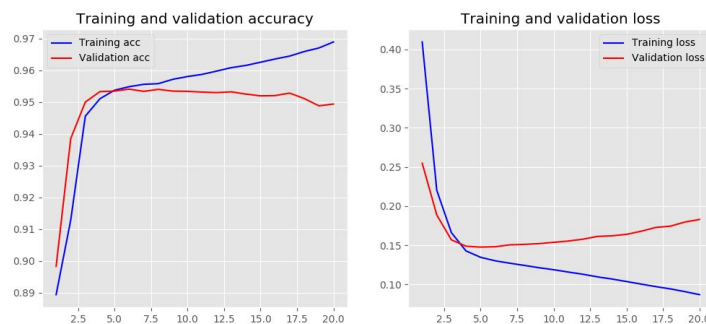
Training Accuracy:  
0.9688  
Validation Accuracy:  
0.9441



Es ist deutlich zu erkennen, dass 50 Iterationen zu viele sind. Das Modell neigt zu 'Overfitting'

Dense: 32  
Iterationen: 20  
Batch Size: 256

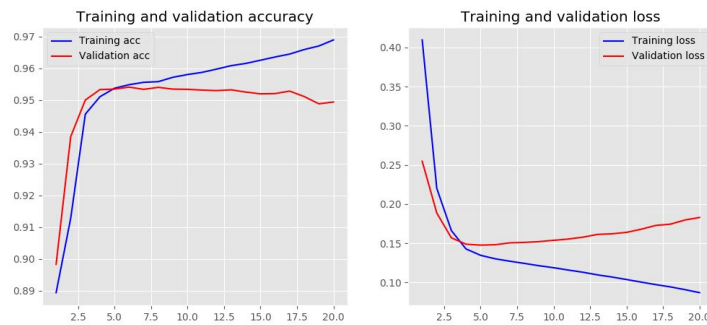
Training Accuracy:  
0.9738  
Validation Accuracy:  
0.9494



Dichte von 32 pro Layer bringt auch keine Verbesserung. Schin nach der fünften Iteration entfernen sich die beiden Kurven voneinander. Am Modell kann leichtes 'Overfitting' beobachtet werden.

Dense: 64  
Iterationen: 20  
Batch Size: 256

Training Accuracy:  
0.9940  
Validation Accuracy:  
0.9450

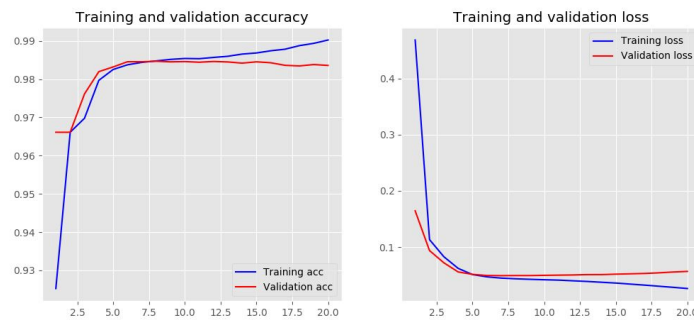


Dito oben. Deutlich zu hohe Dichte pro Layer.

Alle oben aufgelistete Modell wurden mit zwei Klassen trainiert. Wenn man jedoch die Testdaten beobachtet, werden diese in 6 Klassen unterteilt. Aus diesem Grund wurde ein neuer Versuch gestartet das Modell mit 6 Klassen zu trainieren.

**Klassen: 6**  
Dense: 16  
Iterationen: 20  
Batch Size: 512

Training Accuracy:  
0.9914  
Validation  
Accuracy:  
0.9836

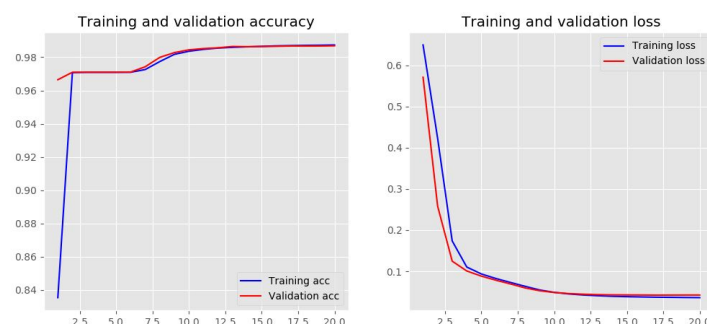


Man kann beobachten, dass die Accuracy sehr gestiegen ist und erreicht um 99%, was ein sehr gutes Ergebnis ist.

Da die 6 Labels sich auf die verschiedenen Klassen von toxischen Kommentaren beziehen braucht man noch eine siebte Klasse für die Gewichtung nicht toxischer Kommentare. Nun kann dieses Modell auch für die Ausführung der Applikation verwendet werden. Resultate im Anhang 'keras 7 classes'

**Klassen : 7**  
Dense: 16  
Iterationen: 20  
Batch Size: 512

Training Accuracy:  
0.9877  
Validation Accuracy:  
0.9868



**Tool:**

- Sequential von keras, TensorFlow as Backend

**Beobachtungen:**

- Die meisten Erkenntnisse wurden oben dokumentiert.
- Für die Applikation mit 2 Klassen wurde das folgende Modell verwendet: Dense: 16, Iterationen: 20, Batch Size: 512 => die Ergebnisse sind sehr ähnlich wie die von Logistic Regression. Die Differenz ist jedoch, dass die nicht eindeutige Kommentare wie 'You are a sad strange little man, and you have my pity.' gefunden werden konnten.
- Bei der Applikation mit 7 Klassen konnte ein interessantes Phänomen beobachtet werden. Die Wahrscheinlichkeitsverteilung konzentriert sich nur auf die ersten beiden Klassen 'neutral' und 'toxic'. Die Gewichtung bei den restlichen Klassen fällt so gering aus, dass diese ignoriert werden kann. Das erklärt, warum das Modell so gut für nur zwei Klassen trainiert werden konnte, siehe 'keras 7 classes'

**Resultate:**

Text-Document: 'hate\_speech\_predict\_keras.txt' oder Anhang 'keras' und 'hate\_speech\_predict\_keras\_6\_classes.txt' oder Anhang 'keras 7 classes'

## SUPPROT VECTOR CLASSIFICATION (SVC) BASELINE

I don't get why negroes always traveling to white countries. Take your ass to Africa!

1

It is a hate comment

I don't get why white girls always traveling to white countries. Go to Africa!

0

It is NOT a hate comment

Only a stupid person considers Google a source of education. I get my facts from real life. You could try getting education from a school. You know what a school is don't you. Its [sic] the place where white people have to pay extra so that 'some people' can get in for a massive discount. Coconuts like you (brown on outside, white on inside) won't survive one day among real black people in Africa.

1

It is a hate comment

What a bunch of bull\$h!+ from an egoistical liberal idiot. I didn't vote for Trump, but I voted. So many people who are mad, didn't. They are world traveling, airbnbers, like this Gen-Y poop, but they're not registered voters. Yet they sure complain when the person they didn't vote for doesn't get elected.

1

It is a hate comment

I miss you. Let's see each other soon!

0

It is NOT a hate comment

I can't stop thinking about you.

1

It is a hate comment

Every time I see you I get butterflies in my tummy.

0

It is NOT a hate comment

You are the light in my life.

0

It is NOT a hate comment

Are you always this stupid or are you making a special effort today

1

It is a hate comment

Calling you an idiot would be an insult to all the stupid people.

1

It is a hate comment

You are a sad strange little man, and you have my pity.  
1  
It is a hate comment

You are my sunshine and I can not live without you  
1  
It is a hate comment

The world is stupid  
1  
It is a hate comment

The world is wonderful  
0  
It is NOT a hate comment

what do you want to know?  
1  
It is a hate comment

it is an absolute shit  
1  
It is a hate comment



## LOGISTIC REGRESSION

I don't get why negroes always traveling to white countries. Take your ass to Africa!

1

It is a hate comment

I don't get why white girls always traveling to white countries. Go to Africa!

0

It is NOT a hate comment

Only a stupid person considers Google a source of education. I get my facts from real life. You could try getting education from a school. You know what a school is don't you. Its [sic] the place where white people have to pay extra so that 'some people' can get in for a massive discount. Coconuts like you (brown on outside, white on inside) won't survive one day among real black people in Africa.

1

It is a hate comment

What a bunch of bull\$h!+ from an egoistical liberal idiot. I didn't vote for Trump, but I voted. So many people who are mad, didn't. They are world traveling, airbnbers, like this Gen-Y poop, but they're not registered voters. Yet they sure complain when the person they didn't vote for doesn't get elected.

1

It is a hate comment

I miss you. Let's see each other soon!

0

It is NOT a hate comment

I can't stop thinking about you.

0

It is NOT a hate comment

Every time I see you I get butterflies in my tummy.

0

It is NOT a hate comment

You are the light in my life.

0

It is NOT a hate comment

Are you always this stupid or are you making a special effort today

1

It is a hate comment

Calling you an idiot would be an insult to all the stupid people.

1

It is a hate comment

Keras: it can not be classified if it is hate comment or not

You are a sad strange little man, and you have my pity.  
0  
It is NOT a hate comment

You are my sunshine and I can not live without you  
0  
It is NOT a hate comment

The world is stupid  
1  
It is a hate comment

The world is wonderful  
0  
It is NOT a hate comment

what do you want to know?  
0  
It is NOT a hate comment

it is an absolute shit  
1  
It is a hate comment

KERAS

I don't get why negroes always traveling to white countries. Take your ass to Africa!

not hate 0.0018118334 hate 0.99871814

It is a hate comment

I don't get why white girls always traveling to white countries. Go to Africa!

not hate 0.91857874 hate 0.08338035

It is NOT a hate comment

Only a stupid person considers Google a source of education. I get my facts from real life. You could try getting education from a school. You know what a school is don't you. Its [sic] the place where white people have to pay extra so that 'some people' can get in for a massive discount. Coconuts like you (brown on outside, white on inside) won't survive one day among real black people in Africa.

not hate 0.06390501 hate 0.93069065

It is a hate comment

What a bunch of bull\$h!+ from an egoistical liberal idiot. I didn't vote for Trump, but I voted. So many people who are mad, didn't. They are world traveling, airbnbers, like this Gen-Y poop, but they're not registered voters. Yet they sure complain when the person they didn't vote for doesn't get elected.

not hate 0.10069539 hate 0.8849401

It is a hate comment

I miss you. Let's see each other soon!

not hate 0.9426023 hate 0.05559118

It is NOT a hate comment

I can't stop thinking about you.

not hate 0.6449997 hate 0.3479226

It is NOT a hate comment

Every time I see you I get butterflies in my tummy.

not hate 0.977434 hate 0.024424898

It is NOT a hate comment

You are the light in my life.

not hate 0.9798075 hate 0.021369899

It is NOT a hate comment

Are you always this stupid or are you making a special effort today

not hate 0.059514306 hate 0.9280194

It is a hate comment

Calling you an idiot would be an insult to all the stupid people.

not hate 0.0002292058 hate 0.999859

It is a hate comment

You are a sad strange little man, and you have my pity.  
not hate 0.55370593 hate 0.45258057  
It is not clear if it's hate comment

You are my sunshine and I can not live without you  
not hate 0.79189074 hate 0.20868471  
It is NOT a hate comment

The world is stupid  
not hate 0.0027705743 hate 0.99794596  
It is a hate comment

The world is wonderful  
not hate 0.9289933 hate 0.07069575  
It is NOT a hate comment

what do you want to know?  
not hate 0.8565512 hate 0.13973302  
It is NOT a hate comment

it is an absolute shit  
not hate 0.0013005944 hate 0.99908507  
It is a hate comment

## KERAS 7 CLASSES

I don't get why negroes always traveling to white countries. Take your ass to Africa!

neutral 0.0013159623  
toxic 0.99875367  
severe\_toxic 0.00083772035  
obscene 0.0009720544  
threat 0.0010536978  
insult 0.002018842  
identity\_hate 0.00094833615  
result: The comment is toxic

I don't get why white girls always traveling to white countries. Go to Africa!

neutral 0.85309714  
toxic 0.13040516  
severe\_toxic 0.0014930221  
obscene 0.005017804  
threat 0.004975261  
insult 0.0021103956  
identity\_hate 0.0031263507  
result: The comment is neutral

Only a stupid person considers Google a source of education. I get my facts from real life. You could try getting education from a school. You know what a school is don't you. Its [sic] the place where white people have to pay extra so that 'some people' can get in for a massive discount. Coconuts like you (brown on outside, white on inside) won't survive one day among real black people in Africa.

neutral 0.07934534  
toxic 0.9121575  
severe\_toxic 9.505993e-06  
obscene 3.1446743e-05  
threat 4.1798423e-05  
insult 2.2632194e-05  
identity\_hate 2.1510717e-05  
result: The comment is toxic

What a bunch of bull\$h!+ from an egoistical liberal idiot. I didn't vote for Trump, but I voted. So many people who are mad, didn't. They are world traveling, airbnbers, like this Gen-Y poop, but they're not registered voters. Yet they sure complain when the person they didn't vote for doesn't get elected.

neutral 0.13509879  
toxic 0.8699682  
severe\_toxic 7.161645e-05  
obscene 0.00023637459  
threat 0.00024436813  
insult 0.00015162911  
identity\_hate 0.00014107156  
result: The comment is toxic

I miss you. Let's see each other soon!

neutral 0.9532005  
toxic 0.044732537  
severe\_toxic 0.00011916373  
obscene 0.000690404  
threat 0.00065129367  
insult 0.00019536541  
identity\_hate 0.00031948308  
result: The comment is neutral

I can't stop thinking about you.

neutral 0.53839445  
toxic 0.4311352  
severe\_toxic 0.0008350565  
obscene 0.0020771655  
threat 0.0022838677  
insult 0.0014460545  
identity\_hate 0.0014666184  
result: The comment is neutral

Every time I see you I get butterflies in my tummy.

neutral 0.9759663  
toxic 0.025680333  
severe\_toxic 0.00052682776  
obscene 0.0026565478  
threat 0.0025812269  
insult 0.00077805785  
identity\_hate 0.0012738942  
result: The comment is neutral

You are the light in my life.

neutral 0.9764009  
toxic 0.024488475  
severe\_toxic 0.00040933836  
obscene 0.0021387108  
threat 0.0020484661  
insult 0.0006090695  
identity\_hate 0.0010087097  
result: The comment is neutral

Are you always this stupid or are you making a special effort today

neutral 0.056268934  
toxic 0.9431206  
severe\_toxic 0.00018425059  
obscene 0.0004656474  
threat 0.00045935257  
insult 0.00038684814  
identity\_hate 0.00030432563  
result: The comment is toxic

Calling you an idiot would be an insult to all the stupid people.

neutral 2.4963876e-05  
toxic 0.99997795  
severe\_toxic 0.0001618429

obscene 0.00021110938  
threat 0.00023529955  
insult 0.00064012356  
identity\_hate 0.00019467795  
result: The comment is toxic

You are a sad strange little man, and you have my pity.  
neutral 0.50186086  
toxic 0.47363746  
severe\_toxic 0.0007238994  
obscene 0.0021362414  
threat 0.0022326654  
insult 0.001191129  
identity\_hate 0.0013952067  
result: The comment is neutral

You are my sunshine and I can not live without you  
neutral 0.7787199  
toxic 0.19989404  
severe\_toxic 0.0008070498  
obscene 0.0025013422  
threat 0.0026031197  
insult 0.0012568382  
identity\_hate 0.0015776701  
result: The comment is neutral

The world is stupid  
neutral 0.0009125067  
toxic 0.99914384  
severe\_toxic 0.00046945285  
obscene 0.00065203867  
threat 0.00064039783  
insult 0.0012430175  
identity\_hate 0.0005185659  
result: The comment is toxic

The world is wonderful  
neutral 0.8553068  
toxic 0.12997608  
severe\_toxic 0.00089293375  
obscene 0.0032995404  
threat 0.0031727147  
insult 0.001378159  
identity\_hate 0.00181871  
result: The comment is neutral

what do you want to know?  
neutral 0.833651  
toxic 0.14483993  
severe\_toxic 0.00021733937  
obscene 0.00079930364  
threat 0.00089723465  
insult 0.00035760368  
identity\_hate 0.0005163117

result: The comment is neutral

it is an absolute shit

neutral 0.00029329845

toxic 0.9997198

severe\_toxic 0.00071540766

obscene 0.00088958396

threat 0.00091135374

insult 0.0021026104

identity\_hate 0.00079175486

result: The comment is toxic