

Massaging the data

...

By Charlie Hill

Twitter Sentiment

	A	B	C	D	E	F	G	H	I	J	
1	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline	name	negat	retweet	text
2	5.70306133677761E+017	neutral	1			Virgin America		cairdin			0 @VirginAmerica What @dh
3	5.70301130888122E+017	positive	0.3486			0 Virgin America		inardino			0 @VirginAmerica plus you've
4	5.70301083672814E+017	neutral	0.6837			Virgin America		yvonnalynn			0 @VirginAmerica I didn't toda
5	5.70301031407624E+017	negative	1	Bad Flight	0.7033	Virgin America		inardino			0 @VirginAmerica it's really a
6	5.70300817074463E+017	negative	1	Can't Tell		1 Virgin America		inardino			0 @VirginAmerica and it's a re
7	5.70300767074181E+017	negative	1	Can't Tell	0.6842	Virgin America		inardino			@VirginAmerica seriously w
8	5.70300616901321E+017	positive	0.6745			0 Virgin America		cjmccinnis			0 it's really the only bad thing
9	5.70300248553349E+017	neutral	0.634			Virgin America		pilot			0 @VirginAmerica yes, nearly
10	5.70299953286943E+017	positive	0.6559			Virgin America		dhepburn			0 @VirginAmerica Really mis
11	5.70295459631264E+017	positive	1			Virgin America		YupitsTate			0 @virginamerica Well, I didn'
12	5.70294189143032E+017	neutral	0.6769			0 Virgin America		idk_but_youtube			0 @VirginAmerica it was ama
13	5.70289724453216E+017	positive	1			Virgin America		HyperCamiLax			0 @VirginAmerica did you kno
14	5.70289584061481E+017	positive	1			Virgin America		HyperCamiLax			0 @VirginAmerica I <
15	5.70287408438121E+017	positive	0.6451			Virgin America		mollanderson			0 @VirginAmerica This is suc
16	5.70285904809599E+017	positive	1			Virgin America		sjespers			0 @VirginAmerica @virginme
17	5.70282469121008E+017	negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon			0 @VirginAmerica Thanks!
18	5.70277724385735E+017	positive	1			Virgin America		ltzBrianHunt			0 @VirginAmerica SFO-PDX
19	5.70276917301137E+017	negative	1	Bad Flight		1 Virgin America		heatherovieda			0 @VirginAmerica So excited
20	5.70270684619923E+017	positive	1			Virgin America		thebrandiray			0 @VirginAmerica I flew from
21	5.70267956648792E+017	positive	1			Virgin America		JNLperce			0 I ♥Flying @VirginAmerica. C
22	5.70265883513385E+017	negative	0.6705	Can't Tell	0.3614	Virgin America		MISSGJ			0 @VirginAmerica you know v
23	5.70264145116819E+017	positive	1			Virgin America		DT_Les			0 @VirginAmerica why are yo
24	5.70259420287869E+017	positive	1			Virgin America		ElvinaBeck			0 @VirginAmerica I love this c
25	5.7025882229758E+017	neutral	1			Virgin America		rjlynch21086			0 @VirginAmerica I love the h
26	5.70256553502069E+017	negative	1	Customer Service Issue	0.3557	Virgin America		ayeevickiee			0 @VirginAmerica will you be
27	5.70249102404923E+017	negative	1	Customer Service Issue		1 Virgin America		Leora13			0 @VirginAmerica you guys n
28	5.70239632807371E+017	negative	1	Can't Tell	0.6614	Virgin America		meredithjlynn			0 @VirginAmerica status mat
29	5.70217831557677E+017	neutral	0.6854			Virgin America		AdamSinger			0 @VirginAmerica What happ
30	5.70207886493782E+017	negative	1	Bad Flight		1 Virgin America		blackjackpro911			0 @VirginAmerica do you mis
31	5.70124596180955E+017	neutral	0.615			0 Virgin America		TenantsUpstairs			0 @VirginAmerica amazing to
32	5.70114021854212E+017	negative	1	Flight Booking Problems		1 Virgin America		jordanpichler			0 @VirginAmerica LAX to EW
											0 @VirginAmerica hi! I just bk

Massaging the data

- Natural Language Toolkit - nltk
- Tokenize, lemmatize and remove stopwords
- X becomes our parsed tweets
- Y becomes our sentiment
- Sklearn's CountVectorizer and Transform methods
- Logistic Regression

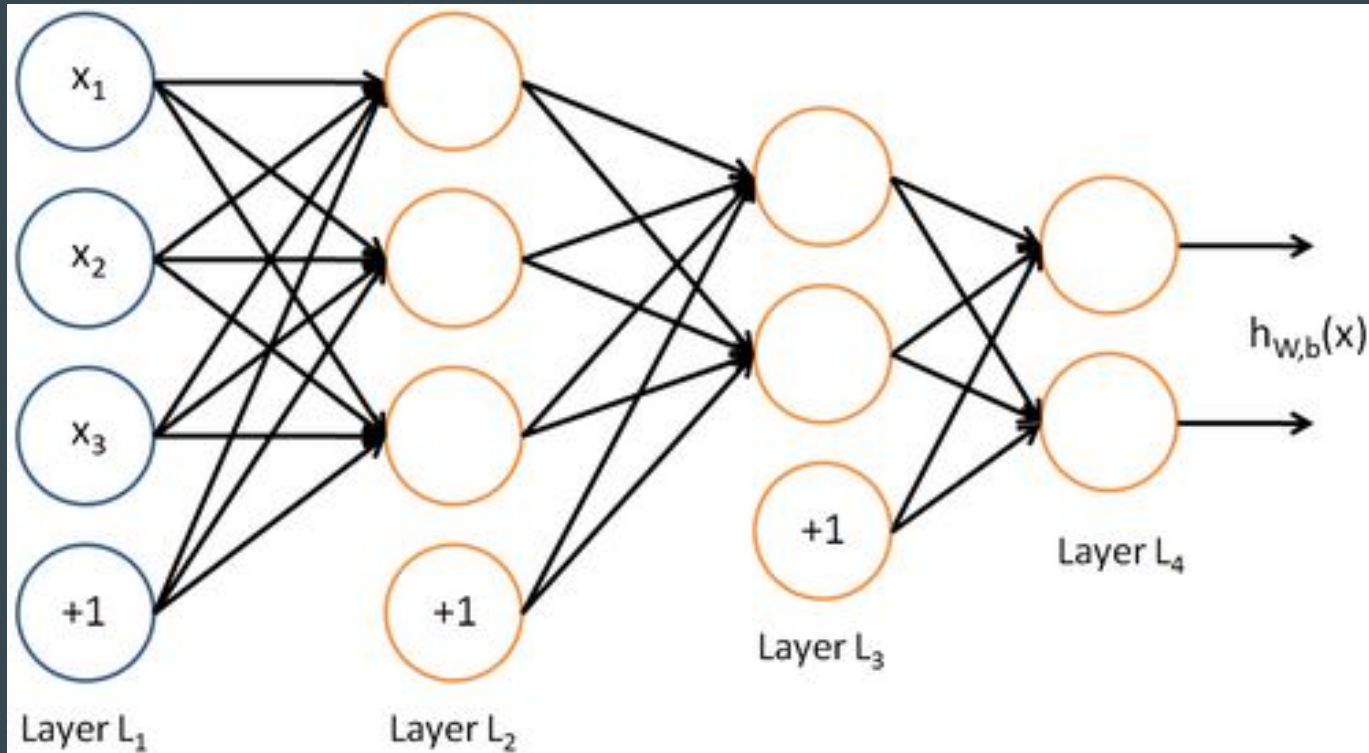
Current Output

```
twl='@united I love this plane'  
print twl  
classify_new_tweet(twl)
```

```
charlie@charlie-VirtualBox:~/Documents/ArtificialIntelligence/Hill/project$ sudo  
python twitterAI.py  
Accuracy : 0.777317880795  
@united I love this plane  
love plane  
Mood of the incoming tweet is: positive  
charlie@charlie-VirtualBox:~/Documents/ArtificialIntelligence/Hill/project$
```

```
charlie@charlie-VirtualBox:~/Documents/ArtificialIntelligence/Hill/project$ sudo  
python twitterAI.py  
Accuracy : 0.777317880795  
@united I hate this plane. Its so sad. Im going to cry  
hate plane sad cri  
Mood of the incoming tweet is: negative  
charlie@charlie-VirtualBox:~/Documents/ArtificialIntelligence/Hill/project$
```

Using a multi output neural network



Working with more simple data

- Kaggle halloween competition

	A	B	C	D	E	F	G
1	id	bone_length	rotting_flesh	hair_length	has_soul	color	type
2	0	0.3545121846	0.350839027	0.4657608918	0.7811416659	clear	Ghoul
3	1	0.575559905	0.425868432	0.5314013787	0.439898877	green	Goblin
4	2	0.4678754987	0.35433042	0.8116160897	0.7912249733	black	Ghoul
5	4	0.7766524607	0.50872255	0.63676558	0.8844636921	black	Ghoul
6	5	0.5661166021	0.875861796	0.4185936709	0.6364378187	green	Ghost
7	7	0.4056797449	0.253277497	0.4414196711	0.2803238199	green	Goblin
8	8	0.3993308509	0.568951767	0.6183910203	0.4679008345	white	Goblin
9	11	0.5162238981	0.536428746	0.6127761466	0.4680482696	clear	Ghoul
10	12	0.3142952759	0.67127969	0.4172669166	0.2275475751	blue	Ghost
11	19	0.2809417441	0.701457493	0.1796332951	0.1411826507	white	Ghost
12	22	0.4316852133	0.438959273	0.239212488	0.4718195627	clear	Goblin
13	23	0.5845432264	0.593081936	0.6811657623	0.9357213331	clear	Ghoul
14	25	0.3907121286	0.335069397	0.5561094618	0.7842165804	white	Ghoul
15	27	0.3515586721	0.471078172	0.4844577817	0.4653278803	black	Goblin
16	28	0.5133872214	0.301344762	0.7456755655	0.545792299	clear	Goblin
17	29	0.500196552	0.438418145	0.5325298148	0.6655224297	clear	Ghoul
18	30	0.2507699998	0.246257686	0.5546544421	0.2500355193	black	Ghost
19	31	0.5855593169	0.585938619	1	0.7086917501	black	Ghoul
20	32	0.6058359561	0.587943179	0.5293612489	0.5267176472	blue	Ghoul
21	34	0.5240803802	0.750988138	0.5246370242	0.4433583431	green	Ghost
22	35	0.5031640578	0.464055334	0.497782803	0.4720310492	clear	Goblin
23	36	0.4726032938	0.427150545	0.5901301906	0.3564882329	white	Ghoul
24	37	0.3744485738	0.384183131	0.6758683608	0.4147003342	clear	Goblin

Massaging the data

- X becomes Bone_length, rotting_flesh, hair_length and has_soul
- 4 dimensions
- Color of the monster appears to be just noise
- Y becomes the type of monster

```
charlie@charlie-VirtualBox:~/Docum
python hw5gendata.py
[sudo] password for charlie:
      _Ghost  _Ghoul  _Goblin
id
0          0.0      1.0      0.0
1          0.0      0.0      1.0
2          0.0      1.0      0.0
4          0.0      1.0      0.0
5          1.0      0.0      0.0
7          0.0      0.0      1.0
8          0.0      0.0      1.0
11         0.0      1.0      0.0
12         1.0      0.0      0.0
19         1.0      0.0      0.0
22         0.0      0.0      1.0
23         0.0      1.0      0.0
25         0.0      1.0      0.0
```

Into the neural network

```
Neurons 10, eta 0.1. Testing set CV score: 0.364006  
Neurons 12, eta 0.1. Testing set CV score: 0.366610  
Neurons 16, eta 0.1. Testing set CV score: 0.370494  
Neurons 18, eta 0.1. Testing set CV score: 0.396582  
Iteration 4, loss = 0.40525050
```

```
Iteration 45, loss = 0.06000000  
Iteration 46, loss = 0.05965803  
Iteration 47, loss = 0.06021860  
Training loss did not improve more than tol=0.000100 for two consecutive epochs. Stopping.  
Training set score: 0.464998  
Testing set score: 0.464998  
charlie@charlie-VirtualBox:~/Documents/ArtificialIntelligence/Hill/project$
```

```
num_hidden = 18  
n_epochs = 1500  
eta = 0.1
```

	A	B
1	id	type
2	3	Ghoul
3	6	Goblin
4	9	Ghoul
5	10	Goblin
6	13	Ghost
7	14	Ghost
8	15	Ghoul
9	16	Ghoul
10	17	Goblin
11	18	Ghoul
12	20	Ghost
13	21	Ghoul
14	24	Goblin
15	26	Ghost
16	33	Ghoul
17	38	Ghost
18	40	Ghoul
19	41	Ghost
20	42	Ghost