Sentiment Classification on UGC data

1. Text Processing

Apply the following procedures on text data.

- All lower-case
- Remove url
- Expand short forms
- Remove punctuations
- Tokenization
- Remove numbers
- Remove stop words

2. Model Selection

2.1 Label Encoding

Encode all labels in train and test set with numbers.

Label	Encoded Label
Negative	0
Neutral	1
Positive	2

2.2 Feature Engineer

Use TF-IDF to create features, ser min_df to 5. The model generates 7015 features. Reduce feature dimension using SVD to 100.

2.3 Modeling and Evaluation

Experiment with 6 models and the results are showed below.

	model_name	accuracy_score	precision_score	recall_score	f1_score
2	Random Forest	0.609339	0.597	0.59635	0.590595
3	AdaBoost	0.566426	0.551304	0.552421	0.545379
5	K Nearest Neighbor	0.54949	0.552702	0.544515	0.540771
1	Decsision Tree	0.479448	0.470986	0.471049	0.470777
4	Gaussian Naive Bayes	0.480434	0.507912	0.476016	0.463437
0	Dummy	0.349227	0.345812	0.346062	0.345819

The best model in terms of overall accuracy is ra ndom forest. Tune the hyperparameter using 5-f old-cross validation and refit the model using be st hyperparameters. The result is showed below.

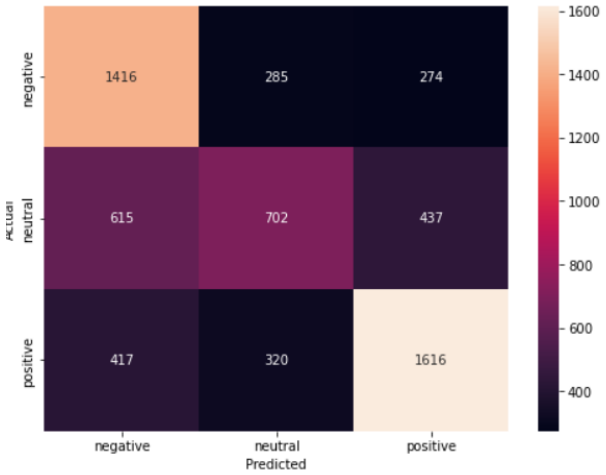
Metric	Score
Accuracy	0.6140
Precision	0.6033
Recall	0.6013
F1	0.5695

3. Result Evaluation

As showed below, positive sentiments are best classified, neutral sentiments are worst classified. There are two potential reasons.

- **Data Imbalance:** the number of positive and negative instances is much larger than the number of neutral instances. This can cause the model to favor the majority class and perform poorly on the minority class.
- Neutral sentiments often **lack clear positive or negative cues**, making them harder to classify.

Label	Precision	Recall
Negative	0.5784	0.7169
Neutral	0.5371	0.4002
Positive	0.6945	0.6868



4. Error Analysis

There are 6 types of errors.

4.1 Negative Classified as Neutral

	cleaned_text	text
35	itchy miserable	I'm itchy and miserable !
41	cant sleep tooth aching	cant sleep . . . my tooth is aching .
48	started think citi really deep gon na survive turmoil gon na next aig	started to think that Citi is in really deep s & ^ t . Are they gonna survive the turmoil or are they gonna be the next AIG ?
77	needs someone explain lambda calculus	needs someone to explain lambda calculus to him ! :(
89	burning cash chrysler gm stop financial tsunami bailout means taking handout	Are YOU burning more cash $ than Chrysler and GM ? Stop the financial tsunami . Where " bailout " means taking a handout !
90	insects infected spinach plant	insects have infected my spinach plant :(
95	history exam studying ugh	History exam studying ugh
137	unfortunate stimulus plan put place twice help gm back american people led inevitable	It's unfortunate that after the Stimulus plan was put in place twice to help GM on the back of the American people has led to the inevitable
160	recovering surgerywishing julesrenner	Recovering from surgery..wishing @julesrenner was here :(
212	naive bayes using em text classification really frustrating	Naive Bayes using EM for Text Classification . Really Frustrating . . .

- **Floppy spelling**: For example, misspell “can't” as “cant”. So the negative "not" is not identified.

- Should not remove some **punctuations** like "!", ':(', ':)' . “!” expresses excitement. ‘:(‘ expresses negative feelings and ‘:)’ expresses positive feelings. Should replace them with their meaning in words.
- The **TF-IDF model** considers the frequency and importance of each word in a corpus, so if the overall context and usage of the words in the corpus do not indicate a negative sentiment, the model may not classify the sentence as negative. This explanation applies to all errors. Data balance is important.

4.2 Negative Classified as Positive

	cleaned_text	text
10	firmly believe obama pelosi zero desire civil charade slogan want destroy conservatism	I firmly believe that Obama / Pelosi have ZERO desire to be civil . It's a charade and a slogan , but they want to destroy conservatism
13	dear nike stop flywire shit waste science ugly love	dear nike , stop with the flywire . that shit is a waste of science . and ugly . love , @vincentx24x
15	talking guy last night telling die hard spurs fan also told hates lebron james	I was talking to this guy last night and he was telling me that he is a die hard Spurs fan . He also told me that he hates LeBron James .
17	lebron beast still cheering atil end	Lebron is a Beast , but I'm still cheering 4 the A. til the end .
32	played android google phone slide screen scares would break fucker fast still prefer iphone	Played with an android google phone . The slide out screen scares me I would break that fucker so fast . Still prefer my iPhone .
34	omg bored tattoooos itchy help aha	omg so bored & my tattoooos are so itchy ! ! help ! aha =)
36	no not itchy maybe later lol	no . I'm not itchy for now . Maybe later , lol .
42	blah blah blah old old no plans today going back sleep guess	Blah , blah , blah same old same old . No plans today , going back to sleep ! guess .
43	glad didnt bay breakers today freaking degrees san francisco wtf	glad i didnt do Bay to Breakers today , it's 1000 freaking degrees in San Francisco wtf
53	annoying new trend internets people picking apart michael lewis malcolm gladwell nobody wants read	annoying new trend on the internets : people picking apart michael lewis and malcolm gladwell . nobody wants to read that .

- Appearance of **positive words** in negative sentence: e.g. firmly believe, dear, cheering, glad
- **Wrong label**: “Lebron is a Beast , but I'm still cheering 4 the A..til the end.” It looks positive to me.

4.3 Neutral Classified as Negative

	cleaned_text	text
69	next time call nike	' Next time , I'll call myself Nike ' .
127	saw new night museum movieit wasokaylol	Just saw the new Night at the Museum movie . .it was . .okay. .lol 7 \ 10
152	time warner cable pulls plug girlfriend experience wwwtinyurlcom	Time Warner Cable Pulls the Plug on ' The Girlfriend Experience ' - (www.tinyurl.com / m595fk)
155	climate focus turns beijing united nations us european governments called china coo	Climate focus turns to Beijing : The United Nations , the US and European governments have called on China to co-o . . .
159	time buy gm car	is now the time to buy a GM car ?
162	dentist tomorrow brush well morning like make hair nice get cut why	Dentist tomorrow . Have to brush well in the morning . Like I make my hair all nice before I get it cut . Why ?
173	waiting line safeway	waiting in line at safeway .
179	jake going safeway	Jake's going to safeway !
180	found safeway picking staples	Found a safeway . Picking up a few staples .
184	normal weight get normal eating blog	Your Normal Weight (and How to Get There) ? Normal Eating Blog

- **Using punctuations without space** is a problem. After removing punctuations, multiple words are joined

4.4 Neutral Classified as Positive

	cleaned_text	text
9	check video president obama white house correspondents dinner	Check this video out - - President Obama at the White House Correspondents ' Dinner
29	need suggestions good ir filter canon got pls dm	need suggestions for a good IR filter for my canon 40D . . . got some ? pls DM
30	checked google business blip shows second entry huh good ba	. I just checked my google for my business - blip shows up as the second entry ! Huh . Is that a good or ba . . . ? – 6emhv
46	san francisco today suggestions	San Francisco today . Any suggestions ?
51	way see star trek esquire	On my way to see Star Trek @ The Esquire .
52	going see star trek soon dad	Going to see star trek soon with my dad .
60	playing curl twitter api	playing with cURL and the Twitter API
62	playing java twitter api	playing with Java and the Twitter API
68	nike owns nba playoffs ads w lebron kobe carmelo	Nike owns NBA Playoffs ads w / LeBron , Kobe , Carmelo ?
70	new blog post nike sb dunk low premium white gum	New blog post : Nike SB Dunk Low Premium ' White Gum '

- Appearance of **positive words** in neutral sentence: e.g. good, play

4.5 Positive Classified as Negative

	cleaned_text	text
4	fair enough think perfect	Fair enough . But i have the Kindle2 and I think it's perfect :)
21	lebron beast nobody nba comes even close	Lebron is a beast . . . nobody in the NBA comes even close .
27	customer innovation award winner booz allen hamilton	[#MLUC09] Customer Innovation Award Winner : Booz Allen Hamilton - -
61	hello twitter api	Hello Twitter API :)
84	work til lets go lakers	work til 6pm . . . lets go lakers ! ! !
110	ever amazing psycop goodby silverstein partners hp go play effects	The ever amazing Psycop and Goodby Silverstein & Partners for HP ! Have to go play with After Effects now !
161	wrist still hurts get looked hate dr dentist scary places time watch eagle eye want join txt	My wrist still hurts . I have to get it looked at . ! HATE the dr / dentist / scary places . :(Time to watch Eagle eye . If you want to join , txt !
165	studing math tomorrow exam dentist	is studing math :) tomorrow exam and dentist :)
197	right lol get high expectations warren buffet style	right ! ! ! LOL we'll get there ! ! I have high expectations , Warren Buffet style .
200	warren buffet became time richest man united states not working investing big idea lead fortune	Warren Buffet became (for a time) the richest man in the United States , not by working but investing in 1 Big idea which lead to the fortune

- **Multiple emotions** within one text
- Should not remove **emoji**
- **Wrong label:** e.g. 165, 200 are hard to tell if they are positive or sarcasm.

4.6 Positive Classified as Neutral

	cleaned_text	text
2	ok first assesment fucking rocks	Ok . first assesment of the #kindle2 . . . it fucking rocks ! ! !
40	going sleep bike ride	is going to sleep then on a bike ride :]
65	scrapbooking nic	is scrapbooking with Nic =D
73	way totally inspired freaky nike commercial	By the way , I'm totally inspired by this freaky Nike commercial .
112	zomg	zomg ! ! ! ! have a G2 ! ! ! ! ! ! !
115	guess retiring start using developer woot googleio	Guess I'll be retiring my G1 and start using my developer G2 woot #googleio
286	highly recommend malcolm gladwell tipping point next audiobook probably one well	I highly recommend Malcolm Gladwell's ' The Tipping Point . ' My next audiobook will probably be one of his as well .
313	learning lambda calculus	Learning about lambda calculus :)
316	atebits finished watching stanford iphone class session really appreciate rock	atebits I just finished watching your Stanford iPhone Class session . I really appreciate it . You Rock !
331	mcdonalds dinner goooood big mac meal	Just had McDonalds for dinner . :D It was goooood . Big Mac Meal . :)

- Should not remove **emoji and exclamations**.

5. Limitation

Did not experiment any deep learning models.

References:

https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f

https://medium.com/@robert.salgado/multiclass-text-classification-from-start-to-finish-f616a8642538

YouTube channel: NormalizedNerd