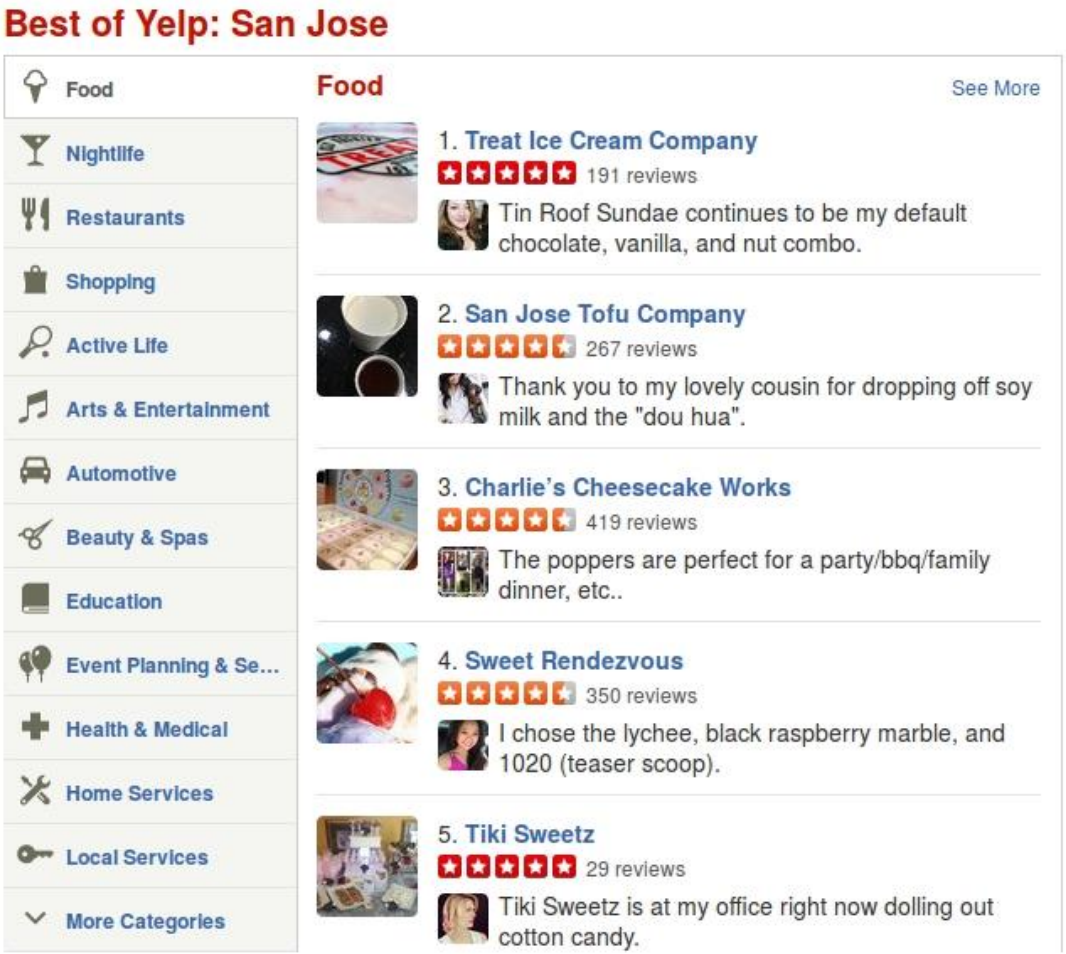
Busines Category Inference from Yelp Data

Bhakti Khude

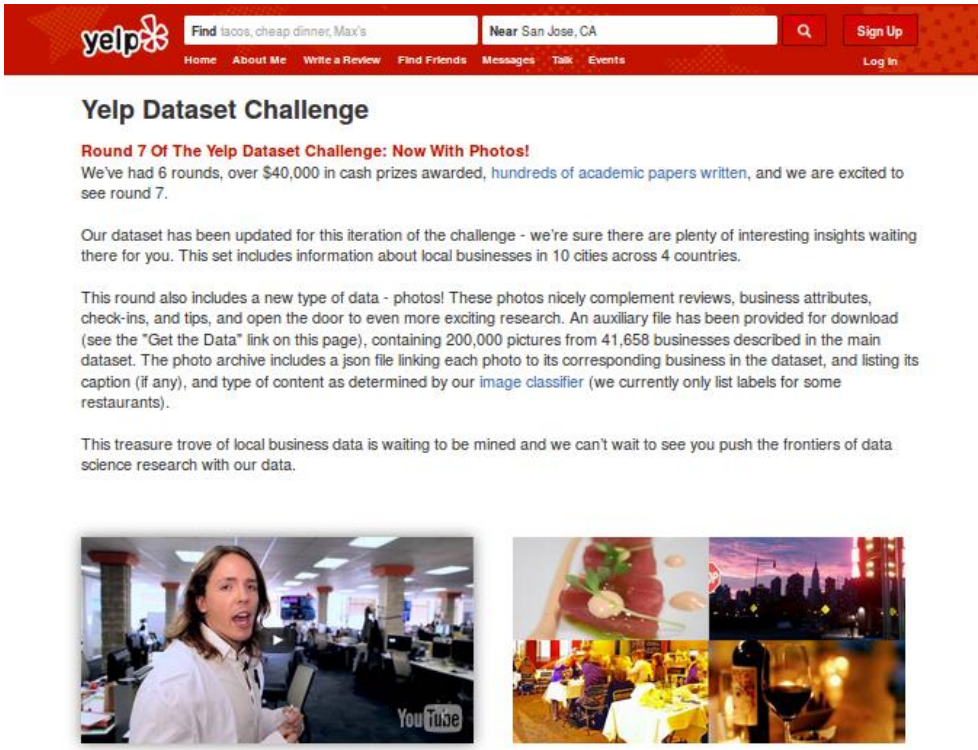
About Yelp

- Yelp.com a crowd-sourced local business review and social networking site

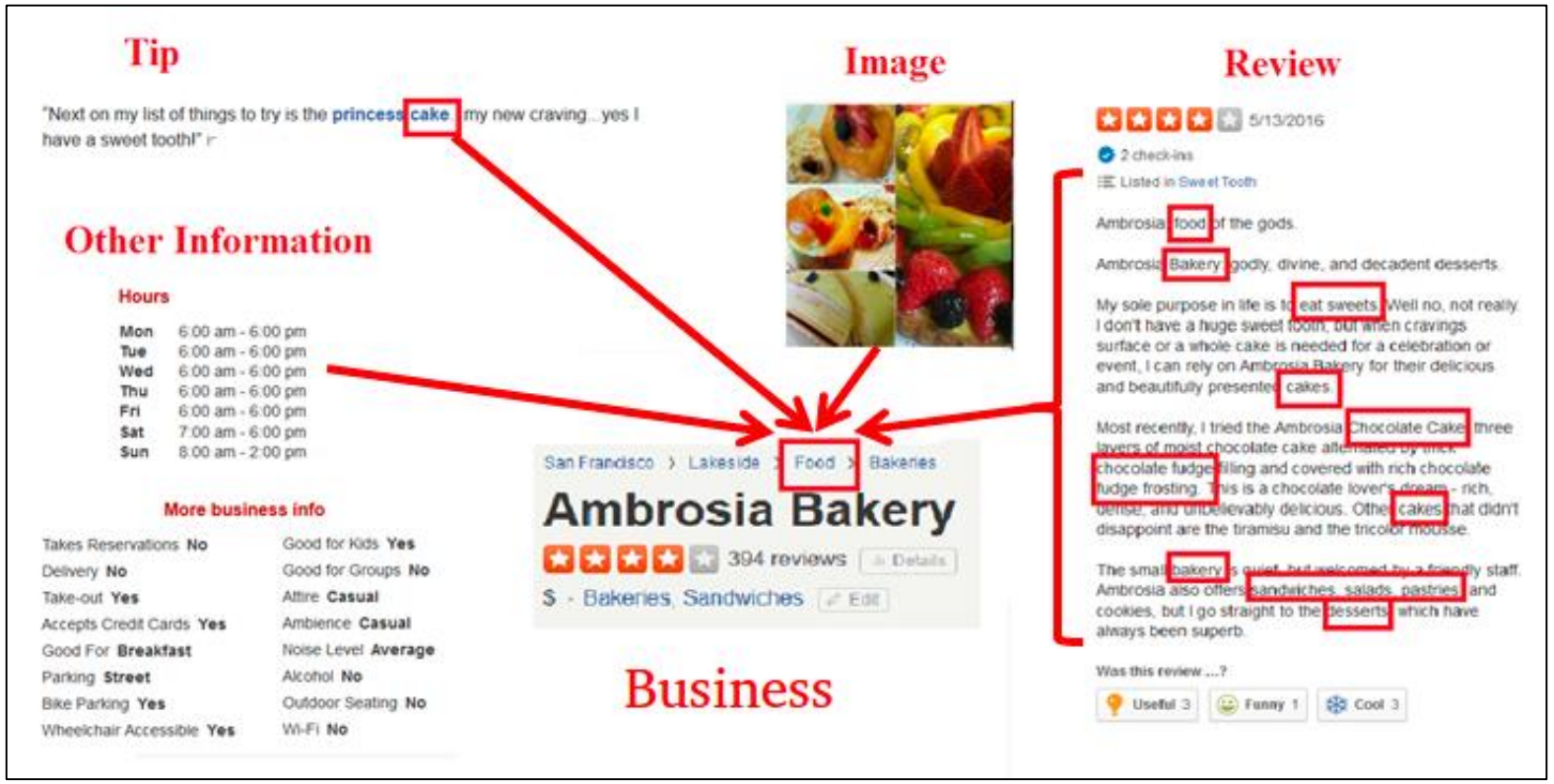


Yelp Dataset Challenge

- 2.2M reviews and 591K tips by 552K users for 77K businesses
- 566K business attributes, e.g. hours, parking availability, ambience
- Social network of 552K users for a total of 3.5M social edges
- Aggregated check-ins over time for each of the 77K businesses
- 2M pictures from the included businesses



Business Classification Model



Build a model using Text and Image data to classify a business

Dataset

Total

- 16,864 Businesses
- 73,576 Reviews
- 36,517 Tips
- 9,026 Images

Training

- 13,500 Businesses
- 59,174 Reviews
- 30,499 Tips
- 6,946 Images

Testing

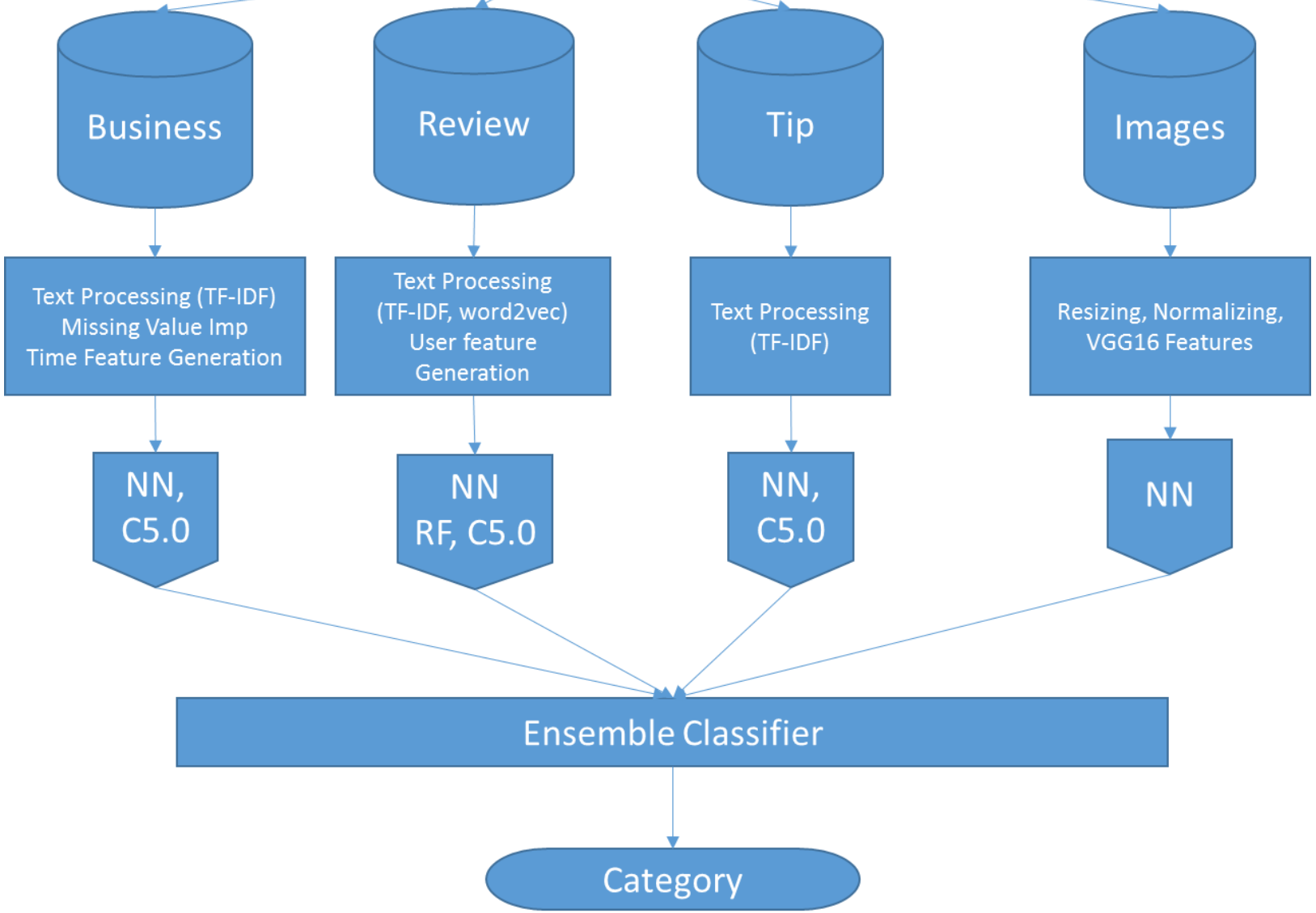
3,364 businesses with atleast a business info & an image. Review and Tip data could be optional.

Data Preprocessing Steps

- Python NLTK package used.
- Foreign accents removed: café was converted to cafe
- Conversion to lowercase, removal of numbers, removal of punctuations, removal of common English “stopwords” (the, is, was etc.)
- Spelling correction
- Lemmatisation
- Parts-of-Speech (POS) Tagging
- Stemming

Modeling

THANK YOU ROB! I truly appreciated all the help I received from this agent today who was able to removed the extra charges on my bill that the Pasadena Verizon Store on Lake was charging me on my bill for upgrading my phone. When I went in I was having problems with my Blackberry and had to switch to the Iphone last week. Rob from the Pennsylvania store who I was connected today was able to look at my bill and all the notes and correct the problem immediately. Great Customer Service! He even set up a FOLLOW UP Phone Call with me On July 5th to make sure the credit goes through on my bill....I can't thank him enough!!!!
After waiting for almost 30 minutes to trade in an old phone part of the buy back program, our customer service rep incorrectly processed the transaction. This led to us waiting another 30 minutes for him to correct it. Don't visit this store if you want pleasant or good service.
I visited this store several months ago to simply ask about smartphone plans. The agent was pleasant and helpful. I would recommend a visit to this store.
Not bad for a Shop N Save, most are privately owned, so availability of items might differ from place to place. The only time I frequent S n S is when the have their huge two day sales usually held on Thursday and Friday at least once month.
Good bakery at this location too.


Business
Review
Tip
Images
Text Processing (TF-IDF)
Missing Value Imp
Time Feature Generation
Text Processing
(TF-IDF, word2vec)
User feature
Generation
Text Processing
(TF-IDF)
Resizing, Normalizing,
VGG16 Features
NN,
C5.0
NN
RF, C5.0
NN,
C5.0
NN
Ensemble Classifier
Category

Modeling

Business

Review

Tip

Images

Text Processing (TF-IDF)
Missing Value Imp
Time Feature Generation

Text Processing
(TF-IDF, word2vec)
User feature
Generation

Text Processing
(TF-IDF)

Resizing, Normalizing,
VGG16 Features

TF-IDF

TF doc-term matrix

	t1	t2	t3	t4	t5	t6
d1	24	21	9	0	0	3
d2	32	10	5	0	3	0
d3	12	16	5	0	0	0
d4	6	7	2	0	0	0
d5	43	31	20	0	3	0
d6	2	0	0	18	7	16
d7	0	0	1	32	12	0
d8	3	0	0	22	4	2
d9	1	0	0	34	27	25
d10	6	0	0	17	4	23

TF-IDF(t1 in D1) = TF*IDF = 24 * log(10/9)

TF-IDF doc-term mat

	t1	t2	t3	t4	t5	t6
d1	2.5	14.6	4.6	0	0	2.1
d2	3.4	6.9	2.6	0	1.1	0
d3	1.3	11.1	2.6	0	0	0
d4	0.6	4.9	1.0	0	0	0
d5	4.5	21.5	10.2	0	1.1	0
...						

IDF weights are (0.1, 0.7, 0.5, 0.7, 0.4, 0.7)

Time feature generation

"Starbucks"

Day	Open	Close
Monday	06:00	21:30
Tuesday	06:00	21:30
Wednesday	06:00	21:30
Thursday	06:00	22:30
Friday	07:00	22:30
Saturday	07:00	21:00
Sunday	06:00	21:30

"Minerva Bakery"

Day	Open	Close
Monday	NA	NA
Tuesday	07:00	16:30
Wednesday	07:00	16:30
Thursday	07:00	16:30
Friday	07:00	16:30
Saturday	07:00	15:30
Sunday	NA	NA

Time slots created: **Morning** : 05:00 - 12:00, **Afternoon** : 12:00 - 17:00, **Evening**: 17:00 - 21:00, **Night**: 21:00 - 05:00

Business	Mor	Aft	Eve	Nit	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Starbucks	1	1	1	1	1	1	1	1	1	1	1
Minerva Bakery	1	1	0	0	0	1	1	1	1	1	0

Category

Modeling

Business

Review

Tip

Images

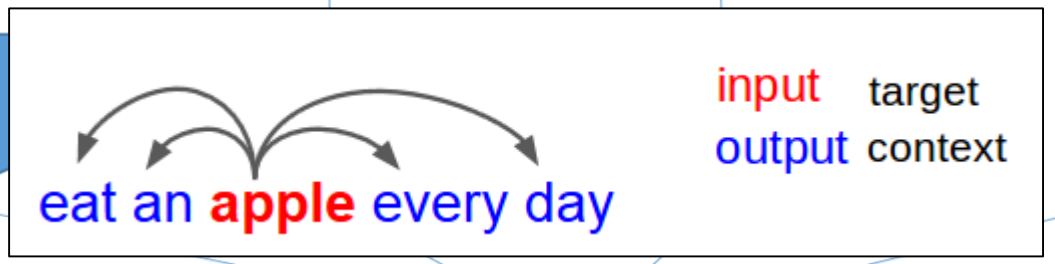
Text Processing (TF-IDF)
Missing Value Imp
Time Feature Generation

Text Processing
(TF-IDF, word2vec)
User feature
Generation

Text Processing
(TF-IDF)

Resizing, Normalizing,
VGG16 Features

NN,
C5.0



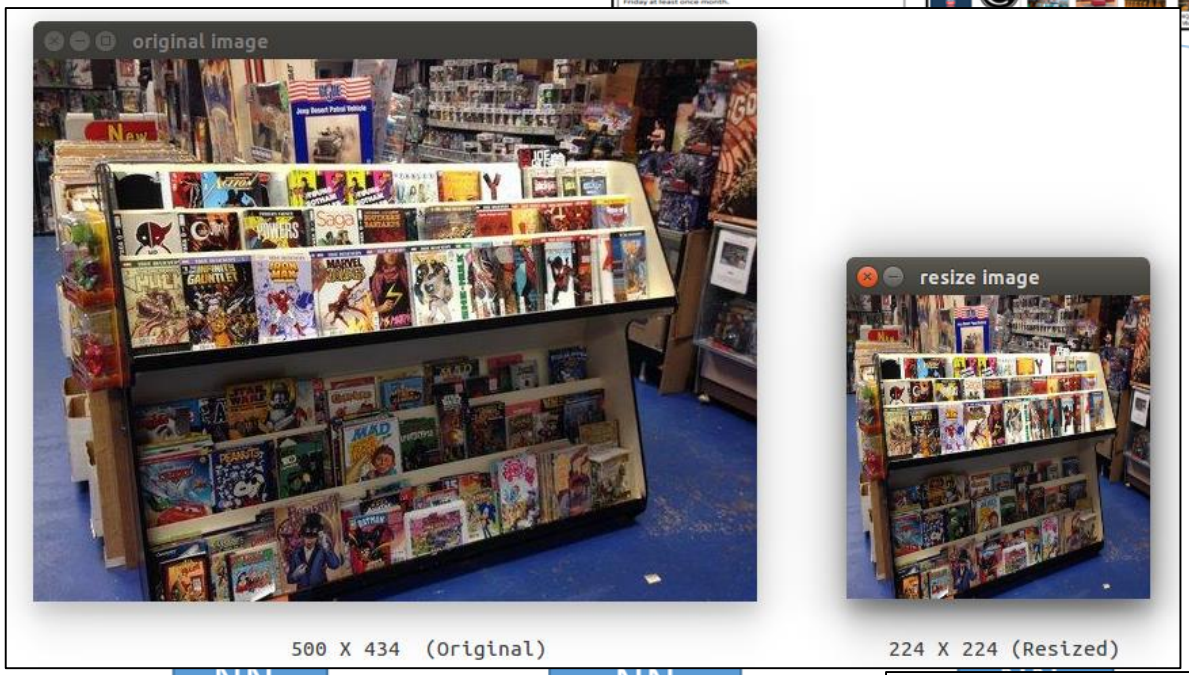
NN

User features

An "Elite" user serves as a local expert.
More the number of years he is Elite more trusted Yelper

User	Elite years	nElite
1	2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015	11
2	-	0
3	2005, 2006, 2007, 2008, 2010, 2011, 2012	7

Modeling



Images

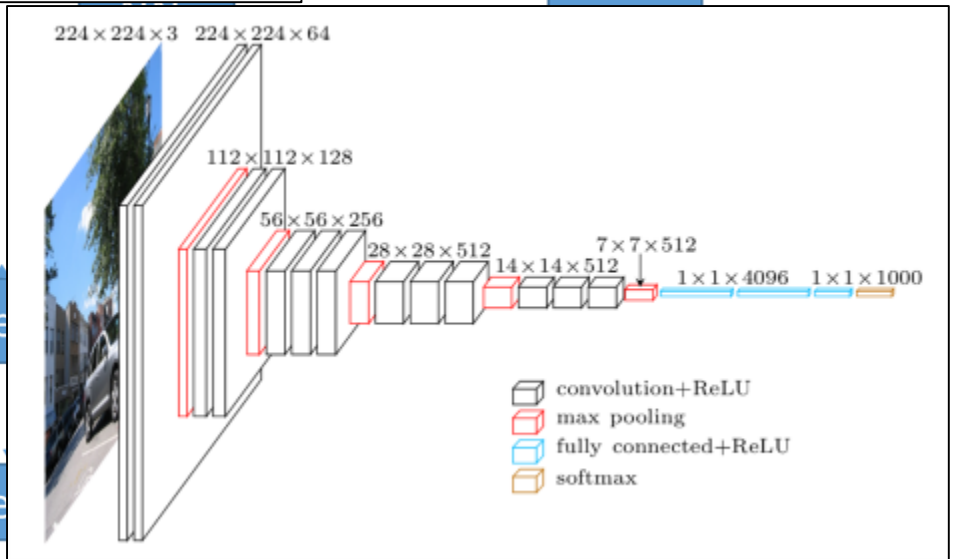
Resizing, Normalizing, VGG16 Features

NN, C5.0

NN RF, C5.0

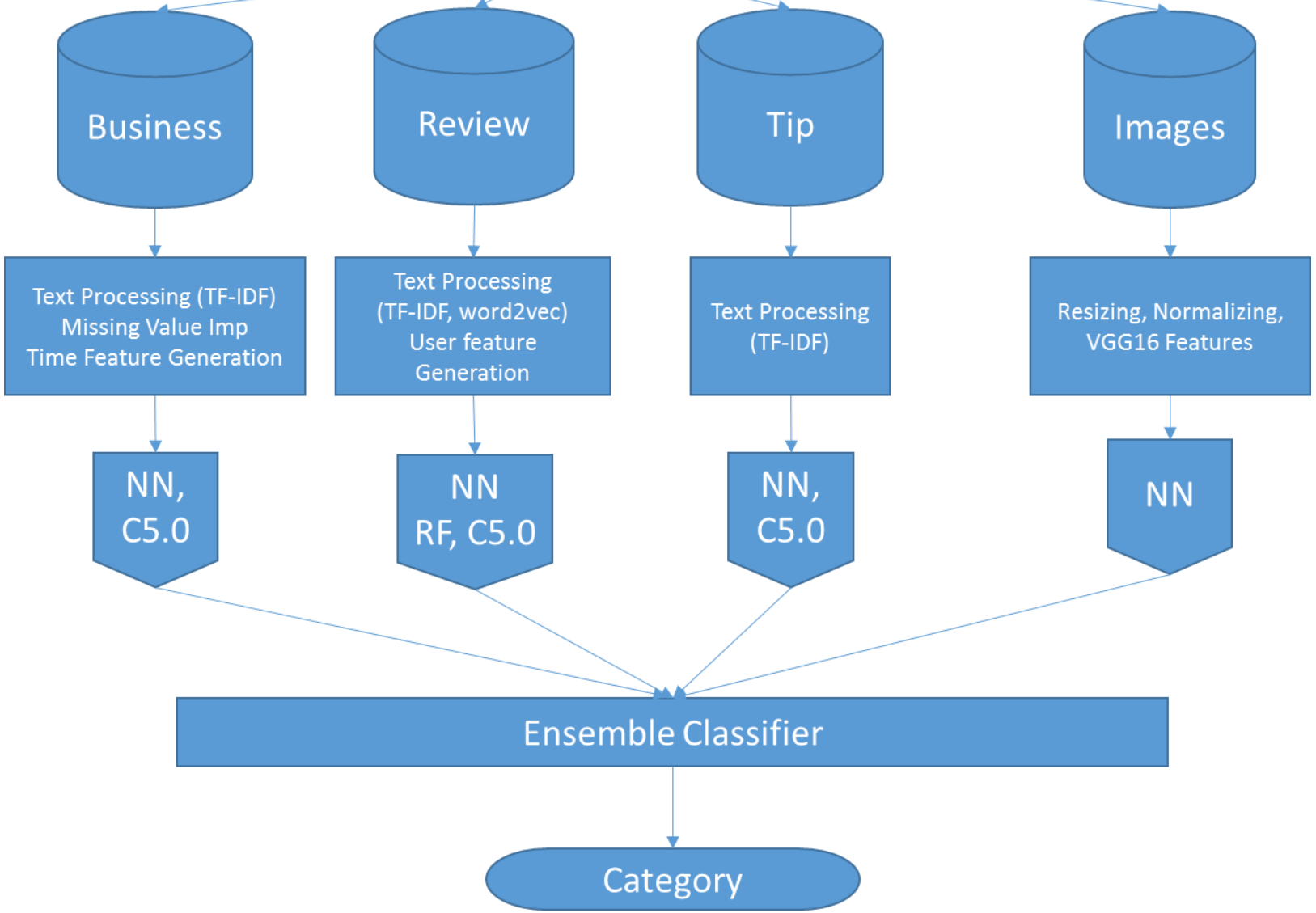
Ensemble

Cate



Modeling

THANK YOU ROB! I truly appreciated all the help I received from this agent today who was able to removed the extra charges on my bill that the Pasadena Verizon Store on Lake was charging me on my bill for upgrading my phone. When I went in I was having problems with my Blackberry and had to switch to the Iphone last week. Rob from the Pennsylvania store who I was connected today was able to look at my bill and all the notes and correct the problem immediately. Great Customer Service! He even set up a FOLLOW UP Phone Call with me On July 5th to make sure the credit goes through on my bill....I can't thank him enough!!!!
After waiting for almost 30 minutes to trade in an old phone part of the buy back program, our customer service rep incorrectly processed the transaction. This led to us waiting another 30 minutes for him to correct it. Don't visit this store if you want pleasant or good service.
I visited this store several months ago to simply ask about smartphone plans. The agent was pleasant and helpful. I would recommend a visit to this store.
Not bad for a Shop N Save, most are privately owned, so availability of items might differ from place to place. The only time I frequent S n S is when the have their huge two day sales usually held on Thursday and Friday at least once month.
Good bakery at this location too.


Business
Review
Tip
Images
Text Processing (TF-IDF)
Missing Value Imp
Time Feature Generation
Text Processing
(TF-IDF, word2vec)
User feature
Generation
Text Processing
(TF-IDF)
Resizing, Normalizing,
VGG16 Features
NN,
C5.0
NN
RF, C5.0
NN,
C5.0
NN
Ensemble Classifier
Category

Results

Data	Model	NEst	Neurons	Epoch	HLoss	Acc	AUC	F1	Prec	Rec	SubAcc
Training											
Bus	NN	-	100,50	300	0.085	0.915	0.915	0.918	0.922	0.914	0.902
Rev	NN	-	100,50	300	0.074	0.926	0.927	0.929	0.937	0.927	0.905
w2v	RF	500	-	-	0.000	1.000	1.000	1.000	1.000	1.000	1.000
Tip	NN	-	100,50	300	0.271	0.729	0.728	0.747	0.743	0.745	0.675
Img	NN	-	4096,1000	100	0.000	1.000	1.000	1.000	1.000	1.000	1.000
ImgN	NN	-	4096,1000	100	0.000	1.000	1.000	1.000	1.000	1.000	1.000
Testing											
Bus	NN	-	100,50	300	0.155	0.845	0.845	0.847	0.849	0.843	0.830
Rev	NN	-	100,50	300	0.068	0.932	0.932	0.933	0.935	0.930	0.919
w2v	RF	500	-	-	0.050	0.950	0.950	0.951	0.949	0.954	0.936
Tip	NN	-	100,50	300	0.328	0.672	0.672	0.681	0.673	0.690	0.642
Img	NN	-	4096,1000	100	0.247	0.753	0.753	0.763	0.745	0.781	0.710
ImgN	NN	-	4096,1000	100	0.219	0.781	0.781	0.790	0.770	0.811	0.735
Ensemble											
1. Tips, No word2vec, Plain Image					0.095	0.906	0.907	0.901	0.960	0.850	0.848
2. No tips, No word2vec, Plain Image					0.072	0.928	0.928	0.929	0.925	0.934	0.908
3. No tips, No word2vec, Normalized Image					0.111	0.889	0.890	0.880	0.979	0.799	0.801
4. No tips, Word2vec, Plain Image					0.054	0.946	0.947	0.945	0.976	0.916	0.916
5. No tips, Word2vec, Normalized Image					0.044	0.958	0.958	0.959	0.961	0.956	0.948