Data Classification on sentiment of Amazon food product review.

Introduction

A gradient of sentiment on any products today is a massive effect to product developers because product designing from inside out may not tailor-made suit to the customers and lead to a total loss, so finding customers insight is probably be the best way to reach the maximum profit in the market shares. A conventional process to find the customers insight is obsolete and lack in both result and time efficiency. In another hand, social media is a massive impact on marketing, it’s modern and simple to comprehend customers feeling.

The mention above is a huge problem and finding the best handling tool is a challenging task. In consequence a comparison in each tool would be analyzed. So, the tools that are powerful, popular, and intensive for sentiment classification now supposed to be K-NN, Naïve bayes, SVM, and Decision tree. The contrast of efficiency on each method may significantly for marketing application by evaluating through free software RapidMiner[1].

This project would be described on a big picture of data and pre-processing to clean the dataset as well, in the first section. Then the dataset would be classified by K-NN, Naïve bayes, SVM, and Decision tree. After that, the result would be evaluated and analyzed through a performance matrix. So, the objectives of this paper are to describe the efficiency of each classifier on Amazon food products and figure out that the cause of negative feedback on products.

Overview

Row No.	ProductId	UserId	Score	Text
180	B005K4Q1VI	ADKBNA7OMK620	negative	It has artificial sweetener which not only taste bad t...
181	B005K4Q1VI	A1Q7232KST2VWE	negative	I have to admit based on most of the reviews I was ...
182	B005K4Q1VI	A24MI2GG040LY7	negative	When I bought my Keurig brewer, I eagerly looked f...
183	B005K4Q1VI	A31IMY00G32AD2	negative	I was hoping this was a good idea. However we tri...
184	B0089SPDUW	A2KZMB70YJARSD	positive	I really enjoy this coffee - I have used a variety of the...
185	B0089SPDUW	A30VFVQ1PX5FTR	positive	Strong smooth velvety flavor. Mahogany describes t...
186	B0089SPDUW	A8OHJUH0WSPVI	positive	Of all the coffees I've tried -- and I've tried A BUNCH ...
187	B0089SPDUW	A107SVKYGPGBPP	positive	If you're a fan of full-bodied coffees, give this one a ...
188	B0089SPDUW	A3JXSK9WWJU7RT	positive	I've sampled many different k-cup coffe types. I pref...
189	B0089SPDUW	AEGL3OL0L6C0B	positive	No bitterness. A strong, full-bodied, cup but does n...
190	B0089SPDUW	AWIU562GSRC76	positive	In my office, and for me personally, this is the smoo...
191	B0089SPDUW	A1U4B3ZM2YVTK9	positive	I have tried about every k cup brand....Caribou Moho...
192	B0089SPDUW	A2VANMJUVLOE4B	positive	If you like a bold cup of good coffee...and have a Ke...
193	B0089SPDUW	A1O299KDNQMPDY	positive	i have this strong flavorful mahogany coffee by carib...
194	B0089SPDUW	A136ONZPRXT76S	positive	We love this coffee. It is the best K cup we've found. ...
195	B0089SPDUW	A2AM935YUELTVQ	positive	My husband enjoyed this coffee. We usually use Ne...

Figure 1 Dataset.

The project focus on sentiment of food product that are being sell on amazon.com in a form of structural dataset in https://www.kaggle.com. The dataset contains a dimension of 11 Features and 525815 objects. However, some features are disadvantages so it must be trim to reduce a size of dataset.

[1] http://rapidminer.com

Moreover, features of dataset is too huge and take time to process on laptop. Thus, the resizing of data set must be done to increase processing performance.

There are multiple attributes in the dataset while a concentration on attribute” Score”, ”Text” in the dataset must be provided at the same time. The “Score” is composed of positive and negative feedback on Amazon food product, another one is “Text” That reflect attitude on products. The attribute “Score” is referred by “Text” Then sentiments has been generated by words that effected satisfactory.

Data pre-processing & Exploration

As mention before, the dimension of dataset must be minimized so Figure 1 represented the dataset that are resized a dimension as 5 features and 819 objects then a summary of sentiment had been generated as shown in figure 3. There are 636 Positive and 183 Negative sentiments. Moreover, a dataset is recognized that is quite cleaned and there are no missing data also redundant on feature “Text”.

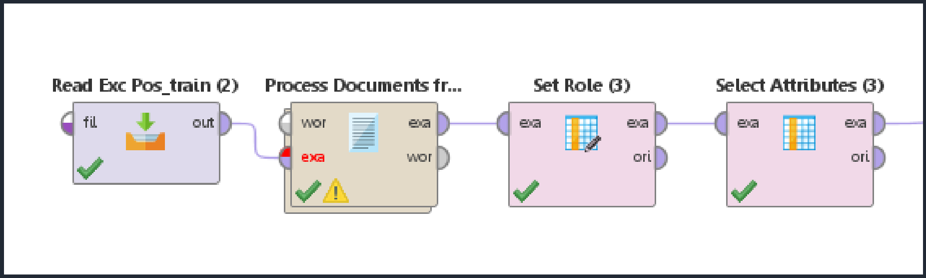


Figure 2 RapidMiner Diagram for data visualization.

The process of parliament plot generated in Figure 3 it would be done by blocks connection of RapidMiner shown in Figure 2 that composed of 2 blocks once “Read Exc Pos_train” and “Process Documents From Data”. The “Read Exc Pos_train_1” block are working by deploying excel file that contained dataset

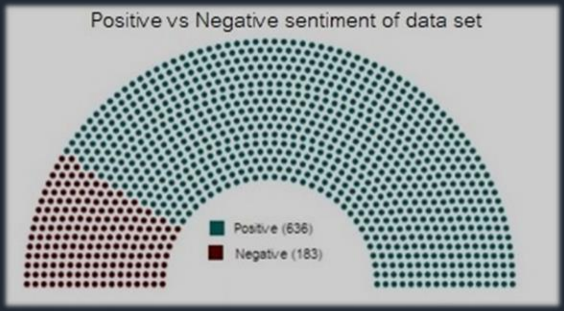


Figure 3 Parliament plot of dataset

Figure 4 Word Cloud on Product ID.

Figure 4 shows the Word Cloud on Product ID. The project focus on 3 most popular sentiment that is “B0016FY6H6”, “B003NDA970” and “B005K4Q1VI” respectively. Thus, generation of bar plot for 3 products by 2 blocks as shown in Figure 5 (Left). Moreover, a majority of sentiment on 3 products is positive feedback in Figure 5 (Right) in the other hand a minority is negative. So, in a business domain, a concentration in majority is probably less benefit in marketing development vice versa focusing on negative feedback can lead a point of view on product improvement by extracting all these negative sentiments and improve a product.

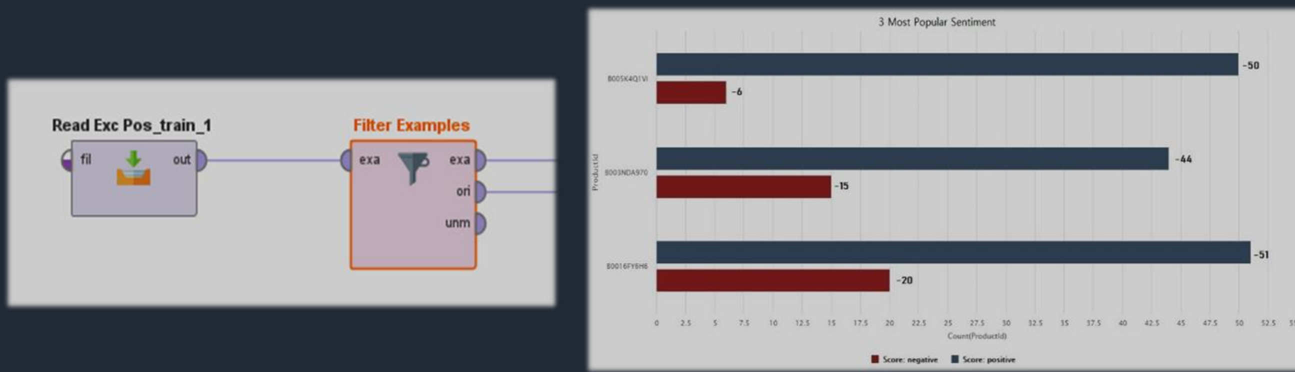


Figure 5 RapidMiner for generating plot (Left), Top 3 sentiment on Amazon product ID (Right)

Modeling & Evaluation

a classification of sentiments would be generated by multiple blocks connection in RapidMiner that composed of multiple usefulness tools to classify each feedback then transform to sentiments that is a comparison of input sentiment and sentiment prediction.

K-NN

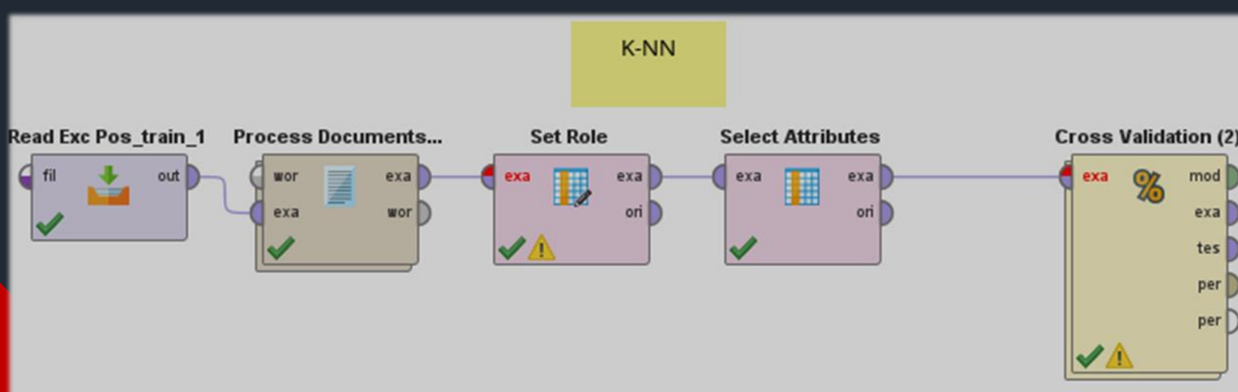


Figure 6 K-NN Process in RapidMiner.

Figure 6 shown the process that started with reading excel file from directory then feedforward to Process Document from data. The inside of this block contained blocks that represent in Figure 7

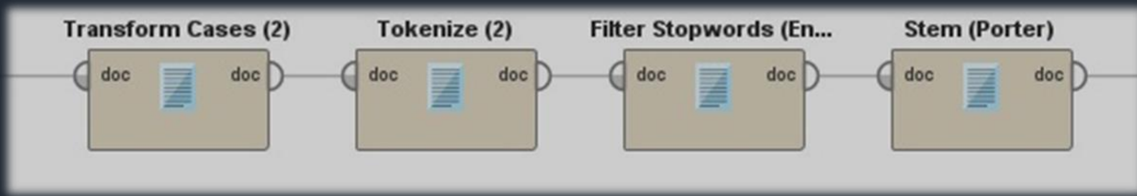


Figure 7 Internal blocks of process document from data both K-NN and Naïve Bayes

accuracy: 78.75% +/- 1.76% (micro average: 78.75%)

	true negative	true positive	class precision
pred. negative	64	55	53.78%
pred. positive	119	581	83.00%
class recall	34.97%	91.35%	

Figure 8 Confusion Matrix of K-NN

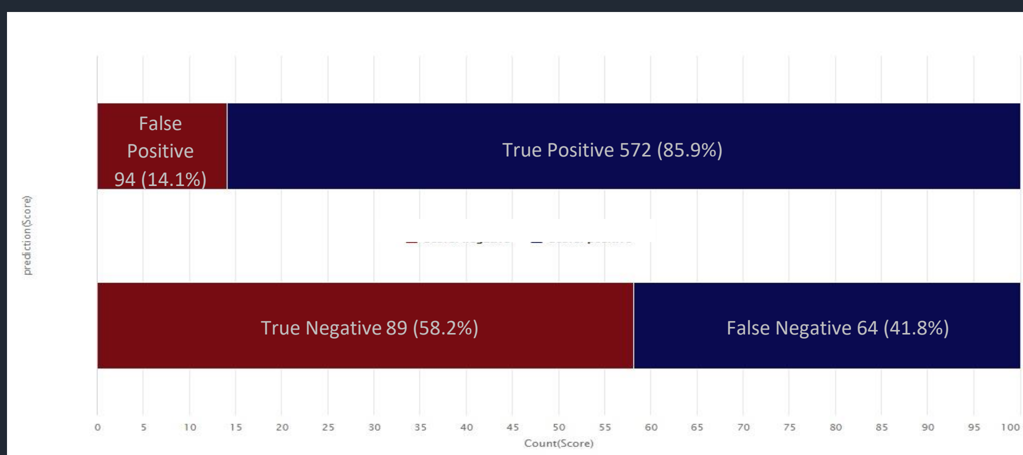


Figure 9 stack to 100% of Confusion Matrix.

calculation

	Calculation	result
Precision	572/(572+94)	0.86
Recall	572/(572+64)	0. 90
F1	2 x (0.90 x 0.86)/(0.90+0.86)	0.87

Naïve Bayes

Another classification algorithm that takes advantage of probability is Naïve Bayes and based on supervised learning theorem[2]. The calculation below represents F1.

2 https://serokell.io/blog/naive-bayes-classifiers

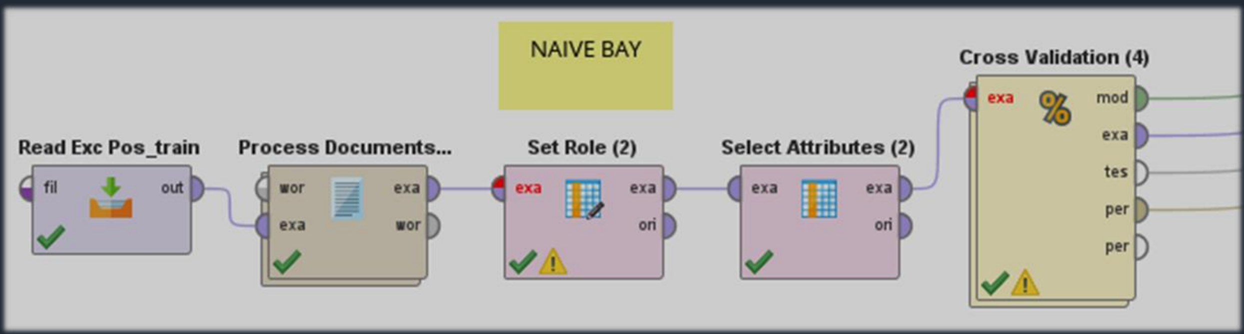


Figure 10 Naïve bayes Process in RapidMiner.

accuracy: 80.70% +/- 5.00% (micro average: 80.71%)

	true negative	true positive	class precision
pred. negative	89	64	58.17%
pred. positive	94	572	85.89%
class recall	48.63%	89.94%	

Figure 11 Confusion Matrix of Naïve Bayes.

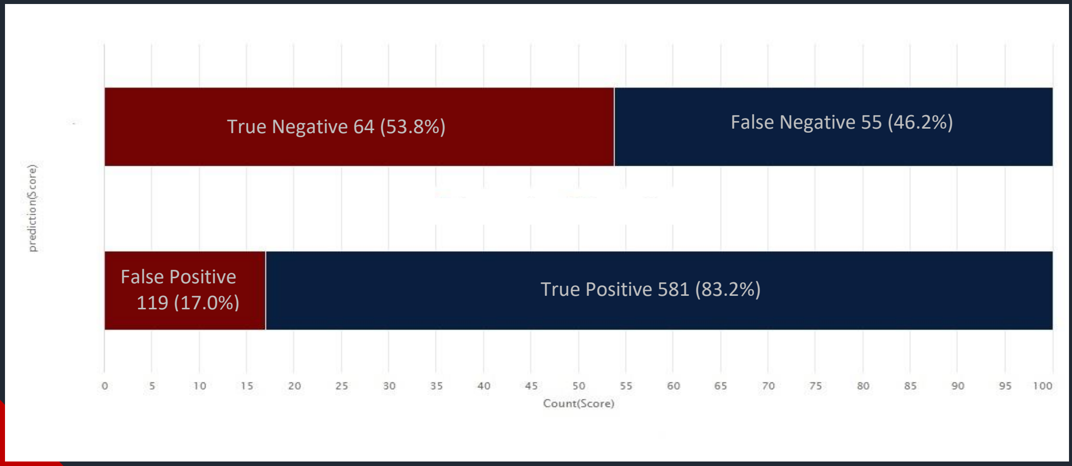


Figure 12 stack to 100% of Confusion Matrix.

The overall process can bring 78.75% of efficiency and 1.76% error. It's unsatisfied to classify a sentiment because False Negative arisen 46.2% that is almost haft. However, the True Positive 83.2% is quite high enough. Then the F1 of the model can be calculated below.

Calculation

	Calculation	result
Precision	581/(581+119)	0.83
Recall	581/(581+55)	0.91
F1	2 x (0.83 x 0.91)/(0.83+0.91)	0.87

Decision Tree

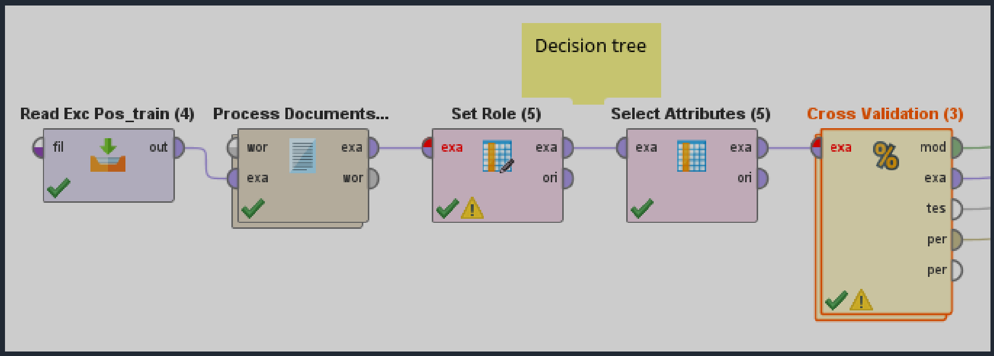


Figure 13 Decision Tree Process in RapidMiner.

Figure 14 Internal blocks of process document from data both decision tree and SVM.

Figure 14 shown the internal process that composed of Transform Cases, it shifted uppercase to lowercase or lowercase to uppercase depending on a setting (in this case need lowercase). Then Tokenize work as a split's operator, it splits text of a document into back of word (tokens). There are several options how to specify the splitting points. Either you may use all non-letter character. This will result in tokens consisting of one single word, what's the most appropriate option before finally building the word vector[3].

After that a filtration of word that unrepresented its meaning in each sentences can be done by Filter Stopwords(ENG). Moreover, Stem(Porter) is the algorithm that applying an iterative, rule-based replacement of word suffixes intending to reduce the length of the words until a minimum length is reached[2]. Finally, the amount of each word can be limited on its word 's characters by Filter Tokens.

[3] RapidMiner help description

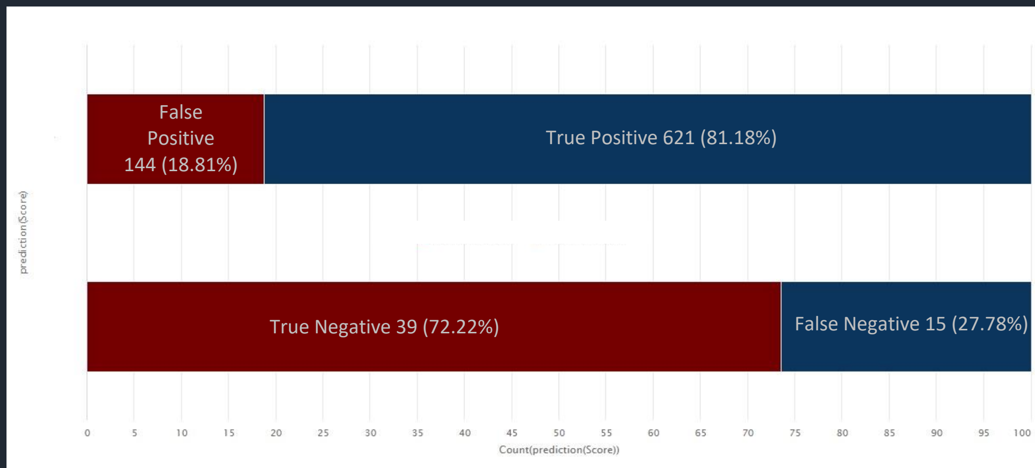
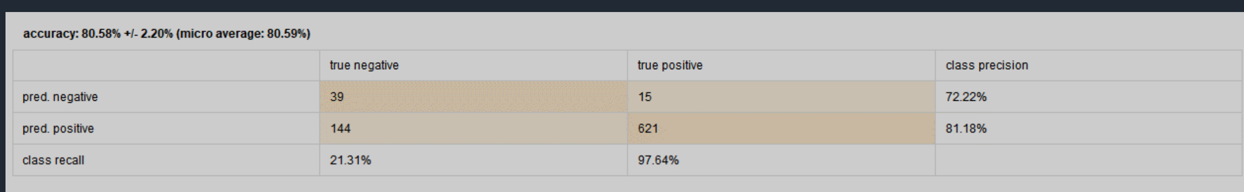


Figure 15 stack to 100% of Confusion Matrix.



accuracy: 80.58% +/- 2.20% (micro average: 80.59%)

	true negative	true positive	class precision
pred. negative	39	15	72.22%
pred. positive	144	621	81.18%
class recall	21.31%	97.64%	

Figure 16 Confusion Matrix of decision tree

calculation

	Calculation	result
Precision	621/(621+15)	0.98
Recall	621/(621+144)	0.81
F1	2 x (0.98 x 0.81)/(0.98+0.81)	0.89

SVM

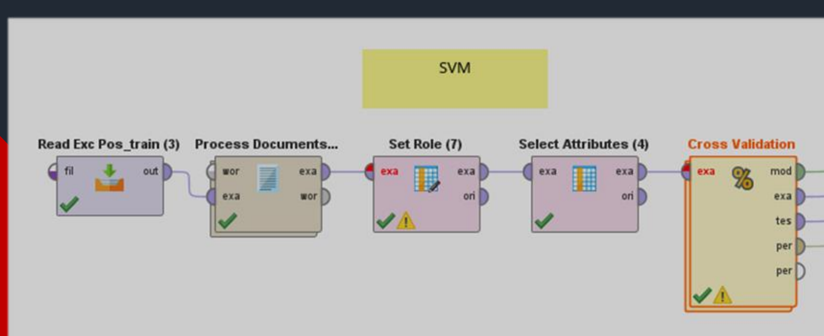


Figure 17 SVM Process in RapidMiner.

accuracy: 84.25% +/- 4.28% (micro average: 84.25%)

	true negative	true positive	class precision
pred. negative	81	27	75.00%
pred. positive	102	609	85.65%
class recall	44.26%	95.75%	

Figure 18 Confusion Matrix of SVM

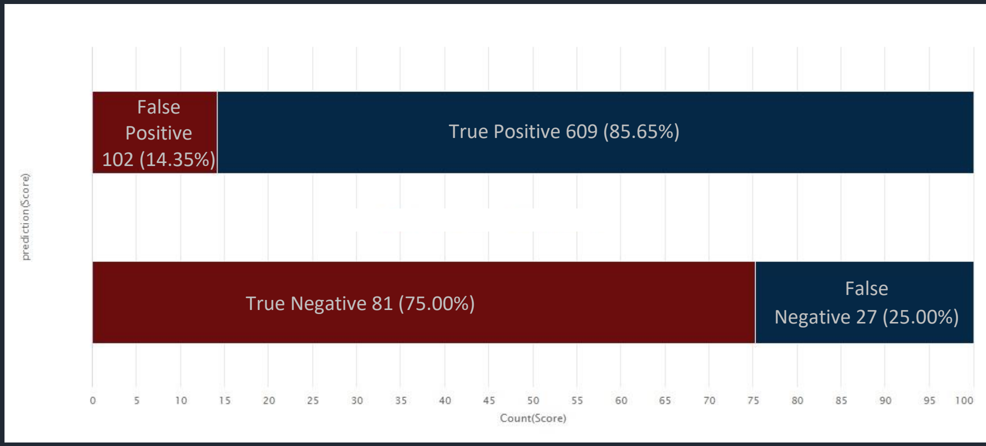


Figure 19 stack to 100% of Confusion Matrix.

Calculation

	Calculation	result
Precision	621/(621+114)	0.86
Recall	609/(609+27)	0.96
F1	2 x (0.96 x 0.86)/(0.96+0.86)	0.91

		K-NN	NAÏVE BAYES	DECISION TREE	SVM
True	Positive	581	572	621	609
True	Negative	64	89	39	81
False	Positive	119	94	114	102
False	Negative	55	64	15	27
Precision		0.83	0.86	0.98	0.86
Recall		0.91	0.90	0.81	0.96
F1		0.87	0.87	0.89	0.91
Accuracy		78.75%	80.70%	80.58%	84.25%

Figure 20 Summary table

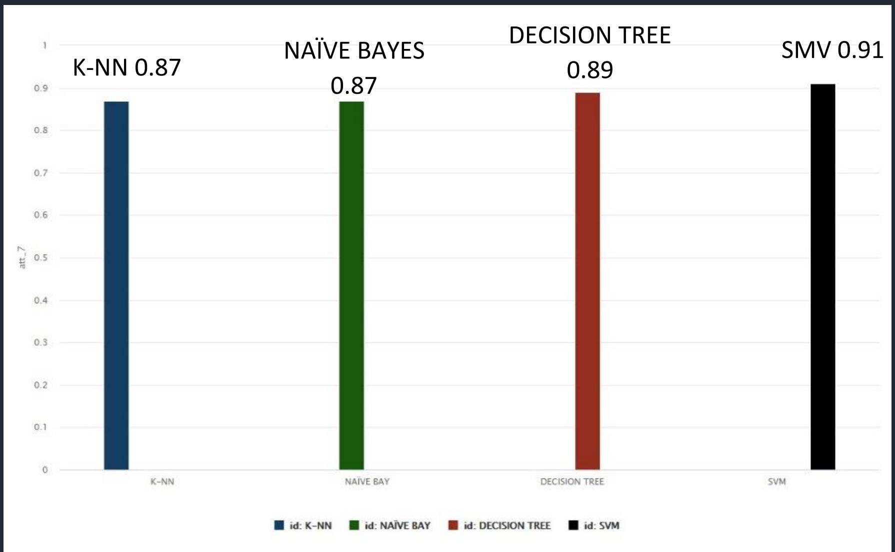


Figure 21 F1

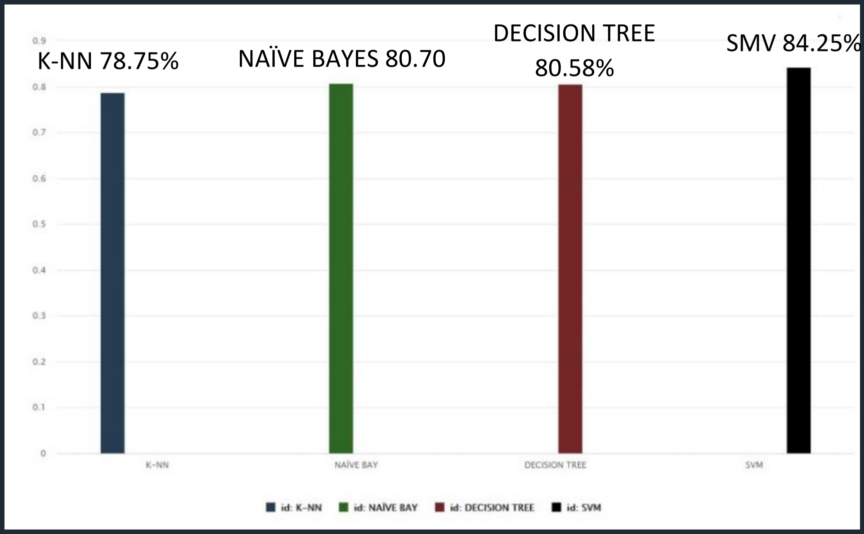


Figure 22 Accuracy

Big 3

Word	Attribut...	Total Occurences ↓	Docum...
popcorn	popcorn	121	52
nda	nda	59	59
pop	pop	57	34
kernel	kernel	28	20
popper	popper	26	16
white	white	23	19
hull	hull	19	13
good	good	17	15
great	great	17	12
ship	ship	17	10
tast	tast	16	15
corn	corn	15	12
love	love	15	14
get	get	14	8
time	time	13	10
flavor	flavor	12	9
oil	oil	12	5
bought	bought	11	10
order	order	11	10
babi	babi	10	9
bag	bag	10	9
lot	lot	10	9

Word	Attribut...	Total Occurences ↓	Document Occurences
tea	tea	180	65
green	green	94	43
water	water	70	35
drink	drink	64	32
product	product	52	29
packet	packet	47	24
flavor	flavor	44	29
powder	powder	40	27
tast	tast	35	27
love	love	31	20
stash	stash	31	21
add	add	30	25
bottl	bottl	28	21
sweeten	sweeten	26	18
mix	mix	25	21
good	good	24	21
great	great	24	19
get	get	23	17
make	make	22	21
sweet	sweet	18	15
lot	lot	17	12
order	order	17	12
sugar	sugar	17	8
look	look	16	11

Word	Attribut...	Total Occurenc... ↓	Document Occurences
hot	hot	51	36
chocol	chocol	49	29
cup	cup	39	27
tast	tast	26	20
cocoa	cocoa	25	18
good	good	20	17
flavor	flavor	17	15
grove	grove	15	11
order	order	15	14
keurig	keurig	14	12
love	love	14	13
product	product	14	11
squar	squar	14	10
coffe	coffe	13	10
tri	tri	13	12
make	make	11	9
milk	milk	11	9
price	price	11	10
great	great	10	9
get	get	9	6
kid	kid	9	8
drink	drink	7	6
pack	pack	7	5
time	time	7	6
try	try	7	5

Figure 23 Word list of top 3 comments, B003NDA970(Left), B0016FY6H6(Middle) and B005K4Q1VI(Right)

The word list shown in Figure 3 a recognition of products by inference of its contexts found that B003NDA970 B0016FY6H6 B005K4Q1VI are popcorn, tea and chocolate

Summary

A comparison of each classification models shown in Figure 20 recognized that there are only positive and negative sentiment so a group must be divided into 2 groups before and the classification model that best classified is SVM because F1 and Accuracy is the effect of evaluation that F1 is the result that reflect the performance of the model, higher F1 reflect high efficiency of the classification model Figure 20 shows that SVM is the highest. Moreover, an accuracy has been verified the efficiency of the model as well.

Thus, to classify a discreate data divided only 2 sentiments(Positive and Negative), The SVM is supposed to be terrific performance more than other. Vice versa, the other 3 classification models provided its performance almost similar.

Moreover, The big 3 in figure 23 represent the common word that are occurred in comments, the most common words in each tables that are noun often represent a kind of the product that sentiment is mention about. as the same time, the context word in the tables always shows result of negative sentiment.

Reference

1. https://www.kaggle.com.
2. https://serokell.io/blog/naive-bayes-classifiers
3. RapidMiner help description