PRASOONA KALLAGUNTA\_HW8

IST 707 \_DATA ANALYTICS

Text Mining - Lie detection

EXPERIMENTING WITH WEKA

Contents

[Goal 3](#_Toc62340244)

[Read the data 3](#_Toc62340245)

[Data Preprocessing 4](#_Toc62340246)

[Support Vector Machine 4](#_Toc62340247)

[Lie 5](#_Toc62340248)

[Sentiment 6](#_Toc62340249)

[Results in the tabular form 6](#_Toc62340250)

[NaïveBayes Multinomial Text 6](#_Toc62340251)

[Sentiment 7](#_Toc62340252)

[Lie 7](#_Toc62340253)

[Results in Tabular Form 7](#_Toc62340254)

[Gain Ratio Attribute Evaluation 8](#_Toc62340255)

[Lie Analysis 8](#_Toc62340256)

[Ranked attributes: 8](#_Toc62340257)

[Sentiment Analysis 10](#_Toc62340258)

[Ranked attributes: 11](#_Toc62340259)

[Chi Squared Attributes 16](#_Toc62340260)

[Lie 16](#_Toc62340261)

[Sentiment 16](#_Toc62340262)

[References 16](#_Toc62340263)

# Goal

As per the instructions provided we must evaluate the claim and figure out whether the customer reviews are fake or true using Support Vector Machine

Here as per the class instructions we can choose either WEKA or Orange. I am choosing to experiment the analysis using WEKA tool.

# Read the data

Data has three columns named as

* + Lie
  + Sentiment
  + Review

i.e., Lie and Sentiment are the nominal values

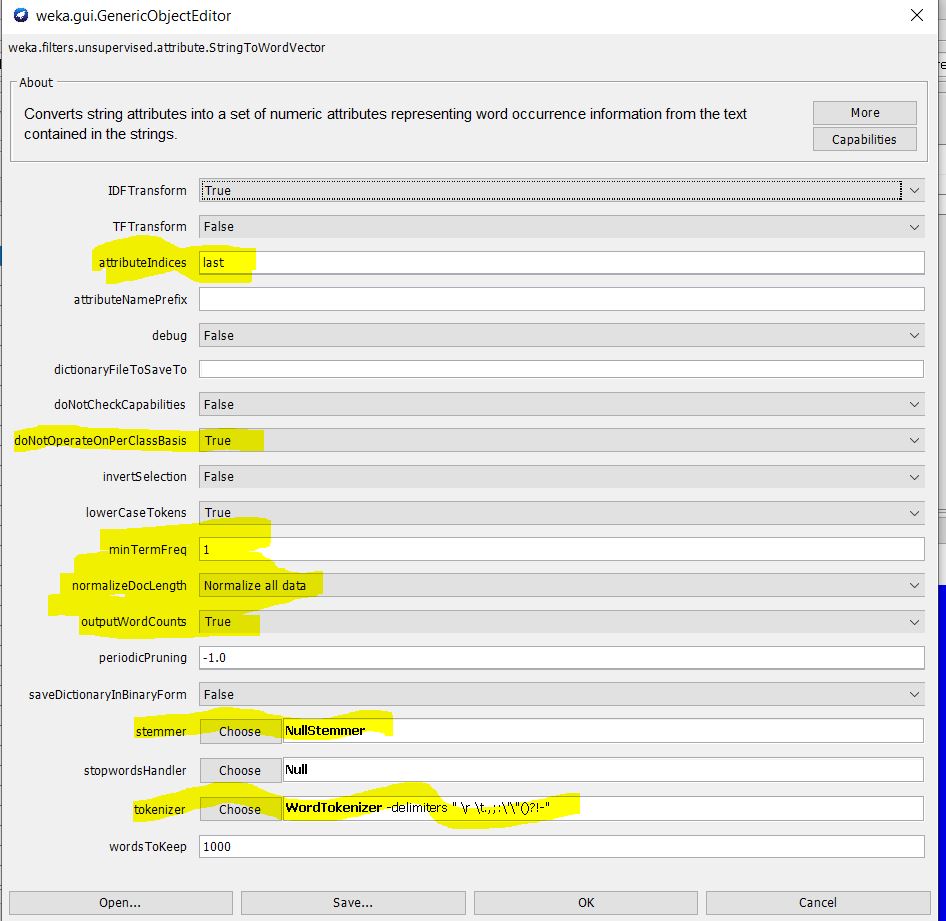
The data can be read in the Weka as “arff” format

# Data Preprocessing

Choose the preprocess tab and select the attributes below is the screen shot

Attributes

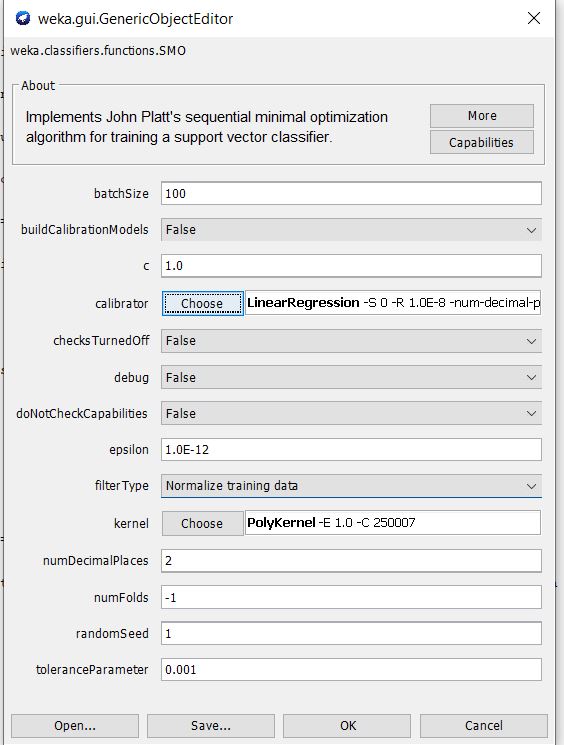
* Filter as String To word Vector
  + Null Stemmer
  + Normalize
  + Turning on IDFTransform
  + Attribute indices to just “last”
  + Add “\_” delimiter
  + Words To Keep as “1000”
  + Min Term Frequency “1”



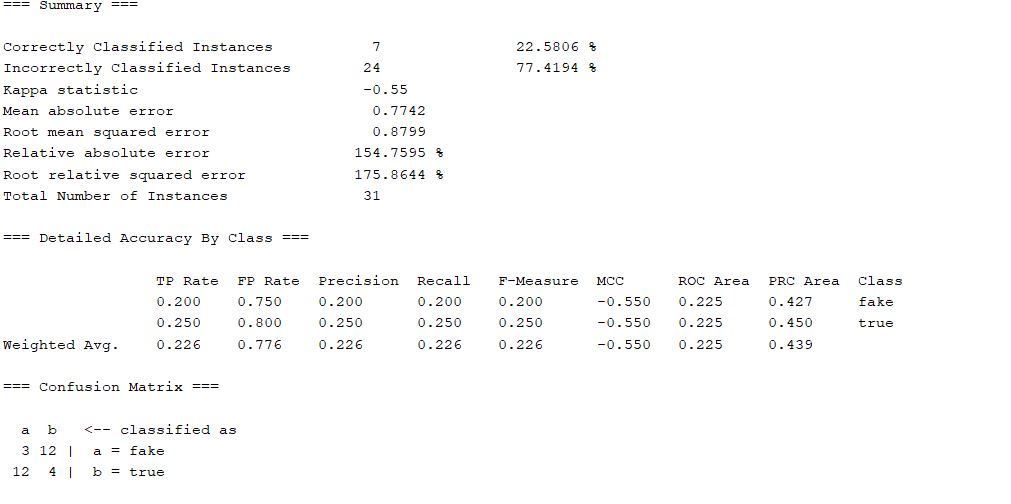
## Support Vector Machine

Using the SVM we can tune the data ensure the following attributes are chosen appropriately

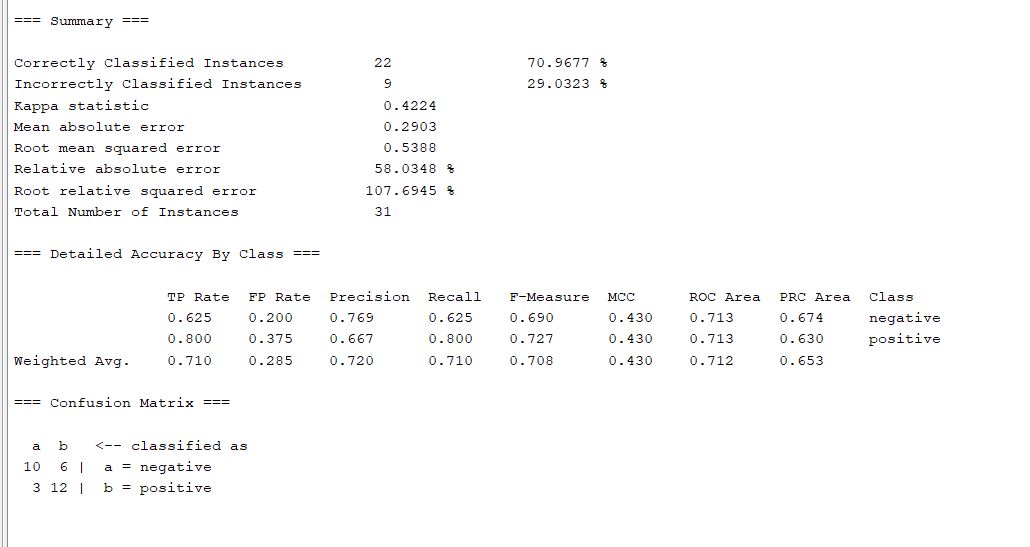
* Callibrator AS “ Linear Regression”



### Lie



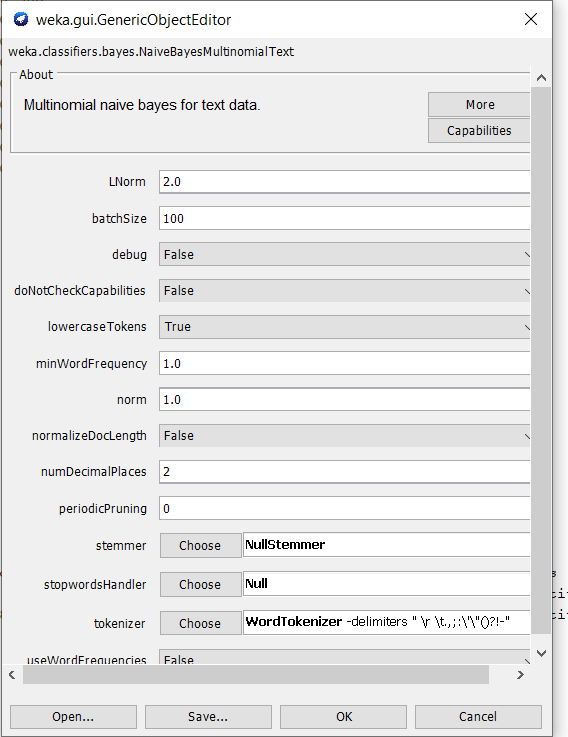
### Sentiment



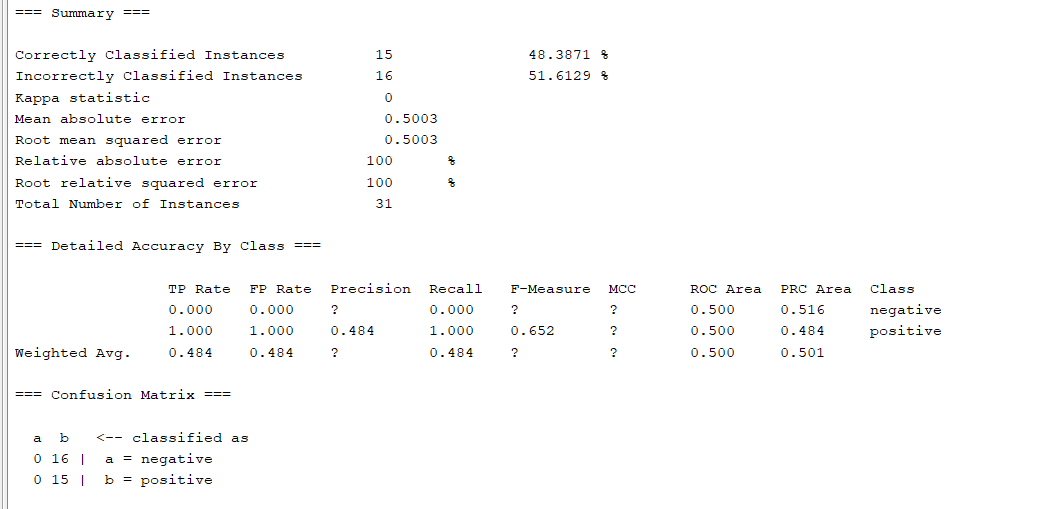
### Results in the tabular form

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter  Setting | Overall  Accuracy | Precision in  Category I | Recall in  Category I | Precision in  Category II | Recall in  Category II |
| Lie | 22.58% | 0.200 | 0.200 | 0.250 | 0.250 |
| Sentiment | 70.96 | 0.769 | 0.625 | 0.667 | 0.800 |

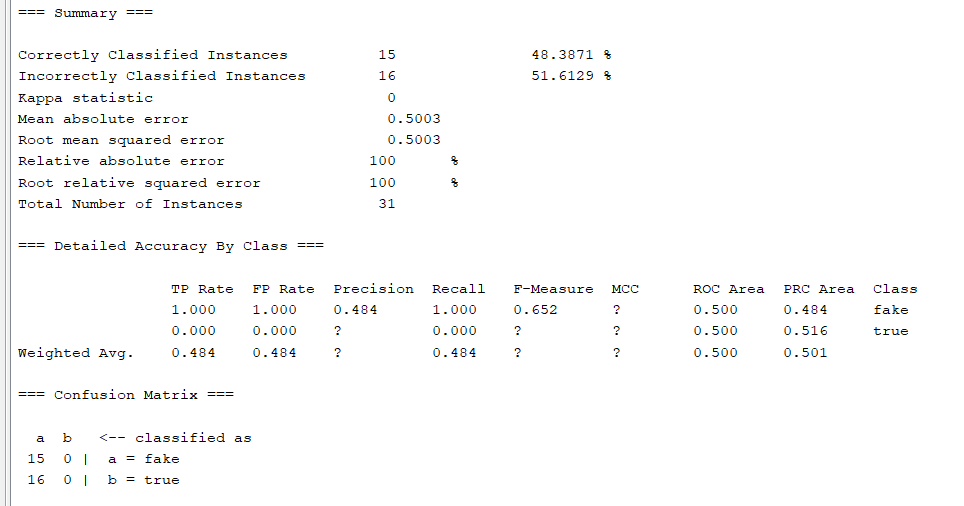
## NaïveBayes Multinomial Text



### Sentiment



### Lie



### Results in Tabular Form

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter  Setting | Overall  Accuracy | Precision in  Category I | Recall in  Category I | Precision in  Category II | Recall in  Category II |
| Lie | 48.38% | 0.484 | 1.000 | ? | 0.000 |
| Sentiment | 48.38% | ? | 0.000 | 0.484 | 1.000 |

# Gain Ratio Attribute Evaluation

## Lie Analysis

Search Method:

Attribute ranking.

**Attribute Evaluator (supervised, Class (nominal): 1 lie):**

Gain Ratio feature evaluator

### Ranked attributes:

0.19694909641295635 566 cold

0.18583265330884222 1002 makes

0.18583265330884222 15 15

0.18583265330884222 606 could

0.17403523705518012 1444 tea

0.17403523705518012 901 iced

0.17403523705518012 899 ice

0.17403523705518012 17 2

0.17403523705518012 701 drinks

0.16111957516930728 1090 only

0.16111957516930728 745 expensive

0.16111957516930728 547 chocolate

0.16111957516930728 735 everyone

0.16111957516930728 1460 thing

0.16111957516930728 465 big

0.16111957516930728 1073 notice

0.16111957516930728 310 Their

0.16111957516930728 56 American

0.16111957516930728 362 add

0.16111957516930728 954 least

0.16111957516930728 723 entered

0.16111957516930728 523 case

0.16111957516930728 1119 packed

0.16111957516930728 1110 outstanding

0.16111957516930728 299 Syracuse

0.16111957516930728 674 dirty

0.16111957516930728 1329 side

0.16111957516930728 1533 variety

0.16111957516930728 888 horrible

0.16111957516930728 941 kitchen

0.16111957516930728 1432 tables

0.16111957516930728 801 fried

0.16111957516930728 515 calling

0.16111957516930728 995 lunch

0.16111957516930728 1425 sushi

0.16111957516930728 422 average

0.16111957516930728 27 4

0.16111957516930728 829 got

0.16111957516930728 20 20

0.12873224578017708 453 been

0.12873224578017528 387 and

0.1010096939859308 1581 what

0 532 chairs

0 587 confirmed

0 589 consisted

0 531 certain

0 530 celebration

0 517 came

0 529 celebrate

0 588 consider

0 586 complimented

0 520 canned

0 516 calmed

0 537 check

0 539 cheerful

0 538 checking

0 585 complimentary

0 534 chance

0 536 cheap

0 535 charged

0 533 champagne

0 590 consistency

0 528 catch

0 522 carmelized

0 596 consuming

0 593 consists

0 594 conspicuously

0 527 casually

0 595 constantly

0 521 cannot

0 524 casino

0 592 consistently

0 519 can

0 525 casino-restaurant

0 518 campus

0 584 completely

0 597 conversations

0 526 cast

0 591 consistent

0 540 cheese

0 577 comfy

0 541 chewy

0 565 coffee

0 579 company

0 558 clearly

0 564 cod

0 563 coconut

0 562 cockroaches

0 561 coast

0 560 client

0 567 coldly

0 568 combination

0 569 combined

## Sentiment Analysis

= Attribute Selection on all input data ===

Search Method:

Attribute ranking.

**Attribute Evaluator (supervised, Class (nominal): 2 sentiment):**

Gain Ratio feature evaluator

### Ranked attributes:

0.2486494050970959 1448 terrible

0.2384853662132655 381 amazing

0.2283069486065256 1486 took

0.21805074922752 379 always

0.21805074922752 409 asked

0.2076350821360398 1066 no

0.2076350821360398 1613 worst

0.2076350821360398 1276 said

0.19694909641295635 571 come

0.18945334590503266 463 best

0.18583265330884222 1505 two

0.18583265330884222 684 dish

0.18583265330884222 473 bland

0.18583265330884222 893 hour

0.18583265330884222 1295 seated

0.18583265330884222 1002 makes

0.17403523705518012 414 atmosphere

0.17403523705518012 1521 until

0.17403523705518012 382 ambiance

0.17403523705518012 916 ingredients

0.17403523705518012 84 But

0.17403523705518012 701 drinks

0.17403523705518012 989 love

0.17403523705518012 1251 restaurants

0.17403523705518012 1550 waited

0.16111957516930728 879 him

0.16111957516930728 1356 soft

0.16111957516930728 993 low

0.16111957516930728 888 horrible

0.16111957516930728 891 hot

0.16111957516930728 708 eat

0.16111957516930728 1120 paid

0.16111957516930728 1255 reviews

0.16111957516930728 1557 waitstaff

0.16111957516930728 1126 pasta

0.16111957516930728 189 Japanese

0.16111957516930728 710 eating

0.16111957516930728 1281 same

0.16111957516930728 962 liked

0.16111957516930728 629 customer

0.16111957516930728 1460 thing

0.16111957516930728 1073 notice

0.16111957516930728 1309 server

0.16111957516930728 156 Great

0.16111957516930728 1425 sushi

0.16111957516930728 1346 smell

0.16111957516930728 954 least

0.16111957516930728 536 cheap

0.16111957516930728 1315 share

0.16111957516930728 657 did

0.16111957516930728 616 cream

0.16111957516930728 499 buffet

0.16111957516930728 1287 saying

0.16111957516930728 515 calling

0.16111957516930728 677 disappointed

0.16111957516930728 674 dirty

0.16111957516930728 1492 town

0.16111957516930728 846 half

0.16111957516930728 547 chocolate

0.16111957516930728 457 began

0.16111957516930728 228 No

0.16111957516930728 829 got

0.16111957516930728 778 five

0.16111957516930728 1148 pick

0.16111957516930728 768 few

0.16111957516930728 1588 while

0.16111957516930728 422 average

0.16111957516930728 20 20

0.16111957516930728 781 flavors

0.16111957516930728 1236 refused

0.14177659917049423 517 came

0.1414666049691737 1029 minutes

0.12873224578017708 1105 our

0.12873224578017708 803 friendly

0.12873224578017708 796 found

0.12100058053208429 1477 to

0.11737706186659164 933 just

0.11519701868755546 313 They

0.11519701868755546 456 before

0.11519701868755546 430 bad

0.09296384205174905 1058 never

0.08389787183168368 1451 that

0.06700814351340925 1072 not

0.06339731947700393 1570 we

0 538 checking

0 540 cheese

0 539 cheerful

0 537 check

0 541 chewy

0 461 benedict

0 462 bento

0 421 authentic

0 426 awkward

0 423 away

0 546 chinses

0 419 attitude

0 420 audience

0 545 chinese

0 542 chicken

0 544 chill

0 543 childhood

0 535 charged

0 534 chance

0 522 carmelized

0 527 casually

0 526 cast

0 525 casino-restaurant

0 524 casino

0 425 awful

0 523 case

0 453 been

0 528 catch

0 460 below

0 529 celebrate

0 533 champagne

0 532 chairs

0 531 certain

0 530 celebration

0 424 awesome

0 418 attentive

0 470 bite

0 464 better

0 568 combination

0 570 combo

0 569 combined

0 567 coldly

0 572 comes

0 566 cold

0 454 beer

0 415 attendant

0 573 comfort

0 564 cod

0 577 comfy

0 579 company

0 578 coming

0 576 comforting

0 413 ate

0 575 comfortably

0 574 comfortable

0 565 coffee

0 563 coconut

0 417 attention

0 552 chop

0 554 city

0 553 chunks

0 551 choosing

0 459 believe

0 550 choose

0 549 choked

0 555 class

0 556 classic

0 416 attended

0 561 coast

0 562 cockroaches

0 458 being

0 560 client

0 557 clean

0 559 clerks

0 558 clearly

0 548 choice

0 521 cannot

0 471 black

0 488 brewed

0 442 basically

0 469 bit

0 487 breast

0 490 bring

0 486 bread

0 485 boxes

0 489 brilliant

0 441 bartender

0 484 boxed

0 492 brown

0 438 bank

0 493 brownie

0 439 bar

0 468 birthday

0 491 brought

0 440 barely

0 443 bathroom

0 483 box

0 580 compared

0 475 blast

0 448 beautiful

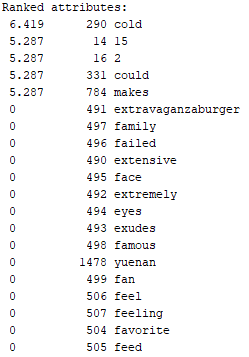
0 449 became

0 474 blanking

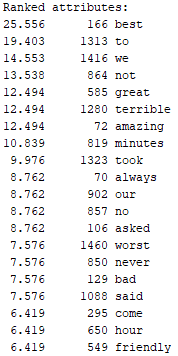
-0.000000000000000891 1 lie

# Chi Squared Attributes

### Lie



### Sentiment



# References

<https://www.tutorialspoint.com/weka/weka_preprocessing_data.htm>

[https://www.tutorialspoint.com/weka/weka\_classifiers.htm](https://www.tutorialspoint.com/weka/weka_classifiers.htm\)