Melisa Kamacı

# CMPE442 – ASSIGNMENT 2 REPORT

**Q1)** The original dictionary that I constructed contained 150138 words. The size of the dictionary is reduced to 10000 words afterward.

**Q2)** Class priors can be seen in Figure 1.

```
Class prior for  0  is:  0.04242531377054976
Class prior for  1  is:  0.05161746508750221
Class prior for  2  is:  0.0522361675799894
Class prior for  3  is:  0.05214778150963408
Class prior for  4  is:  0.05108714866537034
Class prior for  5  is:  0.05241293972070002
Class prior for  6  is:  0.05170585115785752
Class prior for  7  is:  0.05250132579105533
Class prior for  8  is:  0.05285487007247658
Class prior for  9  is:  0.052766484002121264
Class prior for  10  is:  0.0530316422131872
Class prior for  11  is:  0.05258971186141064
Class prior for  12  is:  0.0522361675799894
Class prior for  13  is:  0.05250132579105533
Class prior for  14  is:  0.05241293972070002
Class prior for  15  is:  0.052943256142831886
Class prior for  16  is:  0.048258794414000356
Class prior for  17  is:  0.04984974368039597
Class prior for  18  is:  0.041099522715220084
Class prior for  19  is:  0.033321548523952624
```

*Figure 1: Class Priors*

**Q3)** The confusion matrix can be seen in Figure 2.

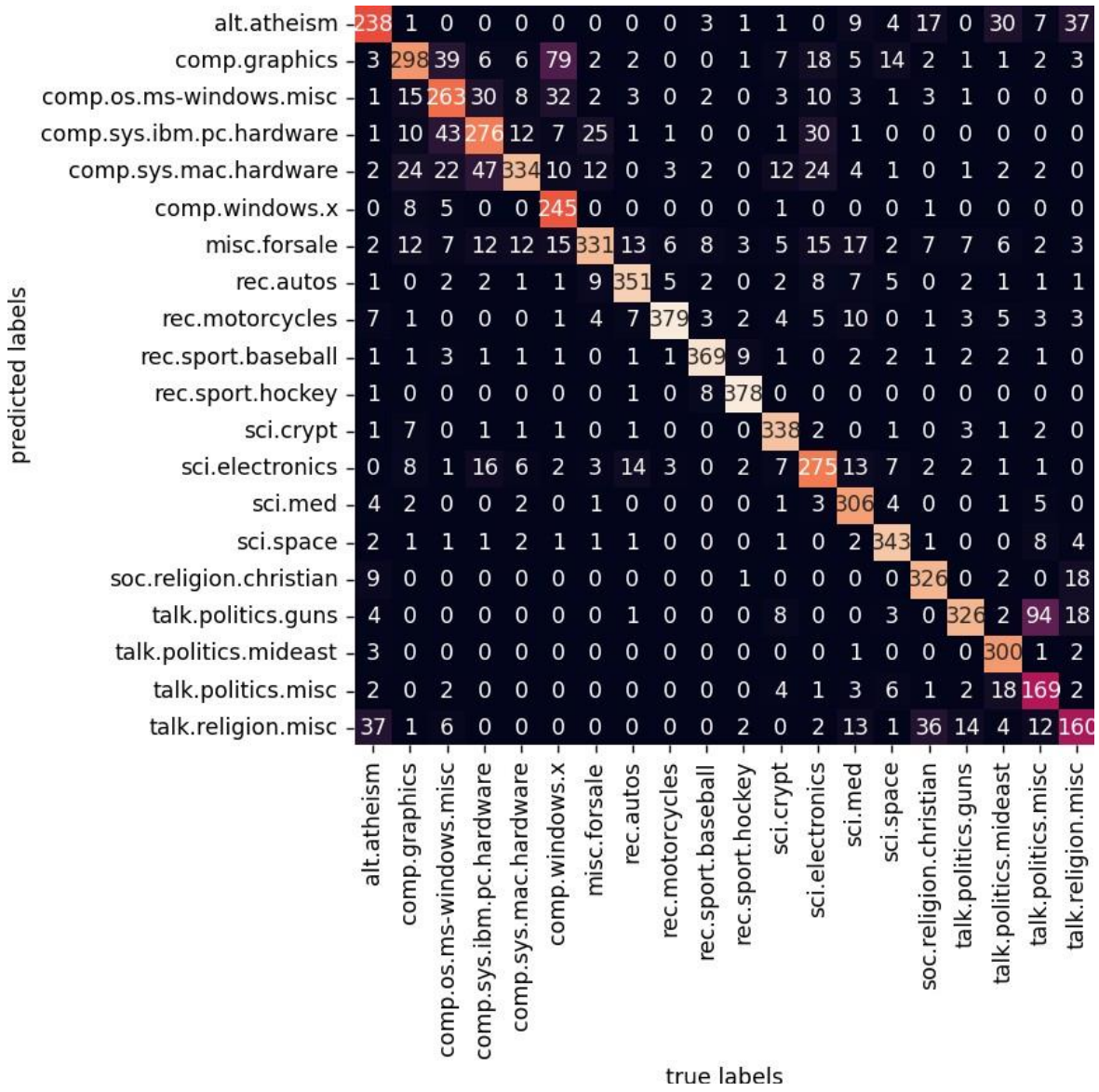


*Figure 2: Confusion Matrix*

Overall testing accuracy can be computed by dividing the summation of left diagonal values by the total number of test documents. The summation of left diagonal values is 6005. Since the total number of test documents is 7532, testing accuracy is equal to $\frac{6005}{7532} = 0.797$. It's important to note that the total computation time of the algorithm is around 7 minutes.

**Q4)** As it can be observed from Figure2, the algorithm sometimes tends to perform inaccurate classifications on "talk.politics.misc" and "talk.religion.misc" labeled documents. "talk.religion.misc" labeled documents are mainly confused as documents of "alt.atheism", "soc.religion.christian" and "talk.politics.guns" categories. "talk.politics.misc" labeled documents are largely misclassified as documents of "talk.politics.guns". These mispredictions are understandable since the misinterpreted topics are closely related to each other.

**Q5)** 10 words that are strong indicators for "alt.atheism" category and their occurrence numbers can be seen in Figure 3.

```
Most commonly occured words in " alt.atheism " documents:
('god', 765)
('one', 722)
('people', 576)
('writes', 562)
('subject', 542)
('line', 538)
('would', 501)
('organization', 472)
('dont', 458)
('atheist', 446)
```

*Figure 3: Most occurred 10 words in "alt.atheism"*

10 words that are strong indicators for “comp.graphics” category and their occurrence numbers can be seen in Figure 4.

```
Most commonly occured words in " comp.graphics " documents:
('line', 774)
('image', 758)
('subject', 627)
('organization', 581)
('file', 554)
('graphic', 460)
('university', 360)
('program', 329)
('would', 295)
('x', 279)
```

*Figure 4: Most occurred 10 words in "comp.graphics"*

10 words that are strong indicators for “comp.os.ms-windows.misc” category and their occurrence numbers can be seen in Figure 5.

```
Most commonly occured words in " comp.os.ms-windows.misc " documents:
('maxaxaxaxaxaxaxaxaxaxaxaxaxax', 3317)
('window', 1082)
('line', 654)
('file', 641)
('subject', 618)
('organization', 579)
('driver', 375)
('university', 333)
('use', 323)
('problem', 318)
```

*Figure 5: Most occurred 10 words in “comp.os.ms-windows.misc”*

10 words that are strong indicators for “comp.sys.ibm.pc.hardware” category and their occurrence numbers can be seen in Figure 6.

```
Most commonly occured words in " comp.sys.ibm.pc.hardware " documents:
('drive', 976)
('scsi', 704)
('line', 637)
('subject', 610)
('organization', 583)
('card', 478)
('system', 410)
('mb', 388)
('one', 379)
('disk', 353)
```

*Figure 6: Most occurred 10 words in "comp.sys.ibm.pc.hardware"*

10 words that are strong indicators for "comp.sys.mac.hardware" category and their occurrence numbers can be seen in Figure 7.

```
Most commonly occured words in " comp.sys.mac.hardware " documents:
('line', 652)
('subject', 591)
('organization', 552)
('mac', 546)
('apple', 420)
('problem', 372)
('drive', 357)
('one', 350)
('university', 308)
('nntppostinghost', 295)
```

*Figure 7: Most occurred 10 words in "comp.sys.mac.hardware"*

10 words that are strong indicators for "comp.windows.x" category and their occurrence numbers can be seen in Figure 8.

```
Most commonly occured words in " comp.windows.x " documents:
('x', 5152)
('window', 933)
('line', 857)
('subject', 788)
('file', 778)
('organization', 599)
('program', 553)
('use', 508)
('widget', 503)
('server', 474)
```

*Figure 8: Most occurred 10 words in "comp.windows.x"*

10 words that are strong indicators for “misc.forsale” category and their occurrence numbers can be seen in Figure 9.

```
Most commonly occured words in " misc.forsale " documents:
('line', 628)
('subject', 601)
('organization', 573)
('sale', 560)
('new', 329)
('university', 326)
('offer', 273)
('nntppostinghost', 263)
('distribution', 252)
('email', 246)
```

*Figure 9: Most occurred 10 words in “misc.forsale”*

10 words that are strong indicators for “rec.autos” category and their occurrence numbers can be seen in Figure 10.

```
Most commonly occured words in " rec.autos " documents:
('car', 1223)
('line', 642)
('subject', 625)
('organization', 589)
('writes', 484)
('article', 453)
('would', 430)
('one', 362)
('like', 327)
('dont', 322)
```

*Figure 10: Most occurred 10 words in “rec.autos”*

10 words that are strong indicators for “rec.motorcycles” category and their occurrence numbers can be seen in Figure 11.

```
Most commonly occured words in " rec.motorcycles " documents:
('bike', 688)
('line', 638)
('organization', 612)
('subject', 611)
('writes', 503)
('article', 473)
('dod', 455)
('one', 394)
('like', 334)
('nntppostinghost', 303)
```

*Figure 11: Most occurred 10 words in “rec.motorcycles”*

10 words that are strong indicators for “rec.sport.baseball” category and their occurrence numbers can be seen in Figure 12.

```
Most commonly occured words in " rec.sport.baseball " documents:
('line', 648)
('subject', 618)
('organization', 601)
('year', 592)
('game', 558)
('writes', 468)
('team', 432)
('article', 391)
('player', 364)
('run', 343)
```

*Figure 12: Most occurred 10 words in “rec.sport.baseball”*

10 words that are strong indicators for “rec.sport.hockey” category and their occurrence numbers can be seen in Figure 13.

```
Most commonly occured words in " rec.sport.hockey " documents:
('team', 954)
('game', 917)
('line', 707)
('subject', 635)
('organization', 611)
('hockey', 594)
('player', 522)
('play', 491)
('year', 438)
('would', 429)
```

*Figure 13: Most occurred 10 words in “rec.sport.hockey”*

10 words that are strong indicators for “sci.crypt” category and their occurrence numbers can be seen in Figure 14.

```
Most commonly occured words in " sci.crypt " documents:
('key', 1449)
('encryption', 840)
('chip', 839)
('would', 707)
('clipper', 705)
('line', 693)
('system', 668)
('subject', 665)
('one', 642)
('organization', 621)
```

*Figure 14: Most occurred 10 words in “sci.crypt”*

10 words that are strong indicators for “sci.electronics” category and their occurrence numbers can be seen in Figure 15.

```
Most commonly occured words in " sci.electronics " documents:
('line', 782)
('subject', 672)
('organization', 581)
('one', 481)
('use', 371)
('would', 367)
('writes', 301)
('university', 291)
('like', 275)
('nntppostinghost', 268)
```

*Figure 15: Most occurred 10 words in “sci.electronics”*

10 words that are strong indicators for “sci.med” category and their occurrence numbers can be seen in Figure 16.

```
Most commonly occured words in " sci.med " documents:
('subject', 649)
('line', 619)
('organization', 610)
('one', 567)
('article', 462)
('writes', 439)
('would', 419)
('people', 322)
('msg', 312)
('dont', 311)
```

*Figure 16: Most occurred 10 words in “sci.med”*

10 words that are strong indicators for “sci.space” category and their occurrence numbers can be seen in Figure 17.

```
Most commonly occured words in " sci.space " documents:
('space', 1200)
('line', 646)
('subject', 635)
('organization', 631)
('would', 553)
('writes', 452)
('one', 413)
('nasa', 407)
('article', 402)
('launch', 377)
```

*Figure 17: Most occurred 10 words in “sci.space”*

10 words that are strong indicators for “soc.religion.christian” category and their occurrence numbers can be seen in Figure 18.

```
Most commonly occured words in " soc.religion.christian " documents:
('god', 1477)
('one', 817)
('would', 775)
('christian', 755)
('subject', 671)
('people', 656)
('line', 636)
('jesus', 618)
('organization', 559)
('say', 520)
```

*Figure 18: Most occurred 10 words in “soc.religion.christian”*

10 words that are strong indicators for “talk.politics.guns” category and their occurrence numbers can be seen in Figure 19.

```
Most commonly occured words in " talk.politics.guns " documents:
('gun', 1231)
('would', 819)
('people', 657)
('line', 608)
('subject', 584)
('organization', 555)
('one', 531)
('writes', 507)
('article', 492)
('right', 487)
```

*Figure 19: Most occurred 10 words in “talk.politics.guns”*

10 words that are strong indicators for “talk.politics.mideast” category and their occurrence numbers can be seen in Figure 20.

```
Most commonly occured words in " talk.politics.mideast " documents:
('armenian', 1268)
('people', 996)
('one', 883)
('israel', 879)
('israeli', 791)
('turkish', 712)
('would', 711)
('subject', 661)
('jew', 630)
('line', 626)
```

*Figure 20: Most occurred 10 words in “talk.politics.mideast”*

10 words that are strong indicators for “talk.politics.misc” category and their occurrence numbers can be seen in Figure 21.

```
Most commonly occured words in " talk.politics.misc " documents:
('would', 692)
('people', 684)
('writes', 611)
('q', 575)
('article', 573)
('line', 525)
('one', 515)
('dont', 508)
('organization', 506)
('subject', 500)
```

*Figure 21: Most occurred 10 words in "talk.politics.misc"*

10 words that are strong indicators for "talk.religion.misc" category and their occurrence numbers can be seen in Figure 22.

```
Most commonly occured words in " talk.religion.misc " documents:
('god', 516)
('one', 463)
('line', 418)
('subject', 418)
('people', 410)
('organization', 403)
('christian', 401)
('jesus', 390)
('would', 387)
('writes', 357)
```

*Figure 22: Most occurred 10 words in "talk.religion.misc"*

**Q6)** Highly related features which lead to redundancy can be removed to achieve a better accuracy rate. By having discrete, powerful indicator features instead of redundant ones, the classifier will be able to make more accurate predictions since it will have more useful tools to work with.