Natural Language Processing

Sentiment Analysis

Submitted to –

Ms. Prabhleen Juneja

Submitted by –

Aman Kumar

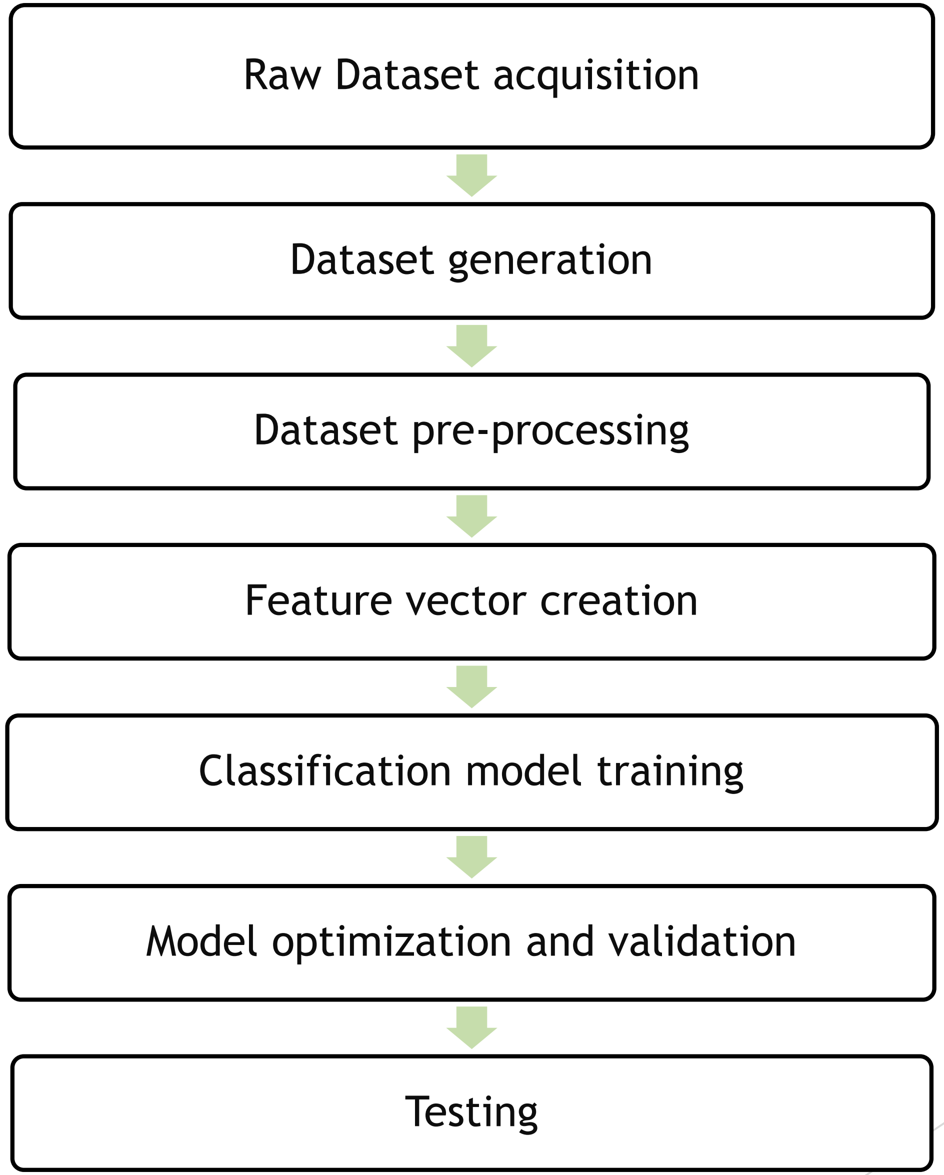
401503006

COE-12

Introduction

- Sentiment analysis is type of text classification
- Instead of classifying a text into a category, here we classify it into a emotion or sentiment for example happy, angry, positive, negative, sad, like, dislike etc
- Why sentiment analysis?
  - Business
  - Politics
  - Public Actions
  - E-Commerce
  - Voice of the Market (VOM)
  - Government

Workflow



Raw Dataset acquisition

Dataset generation

Dataset pre-processing

Feature vector creation

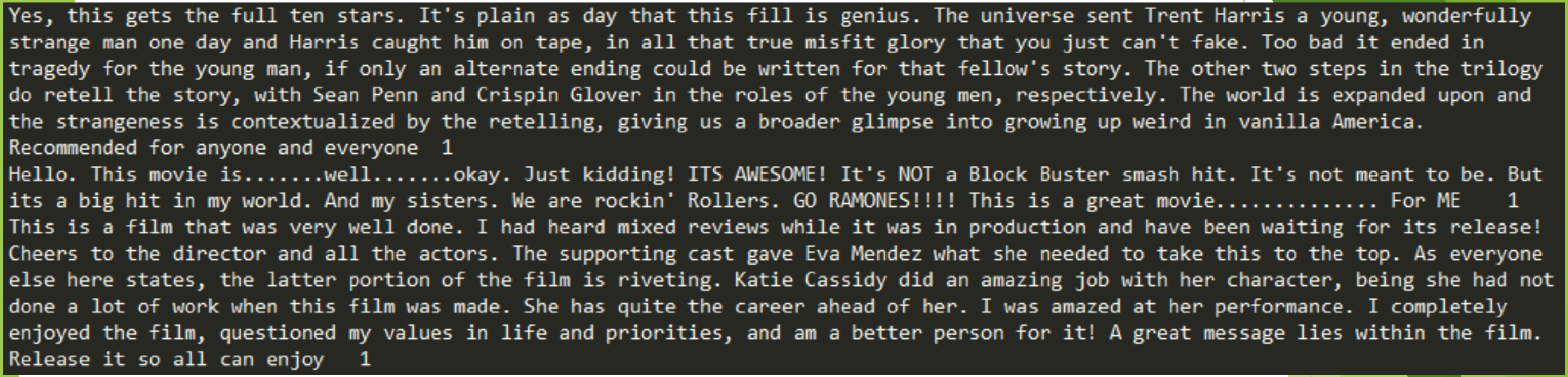
Classification model training

Model optimization and validation

Testing

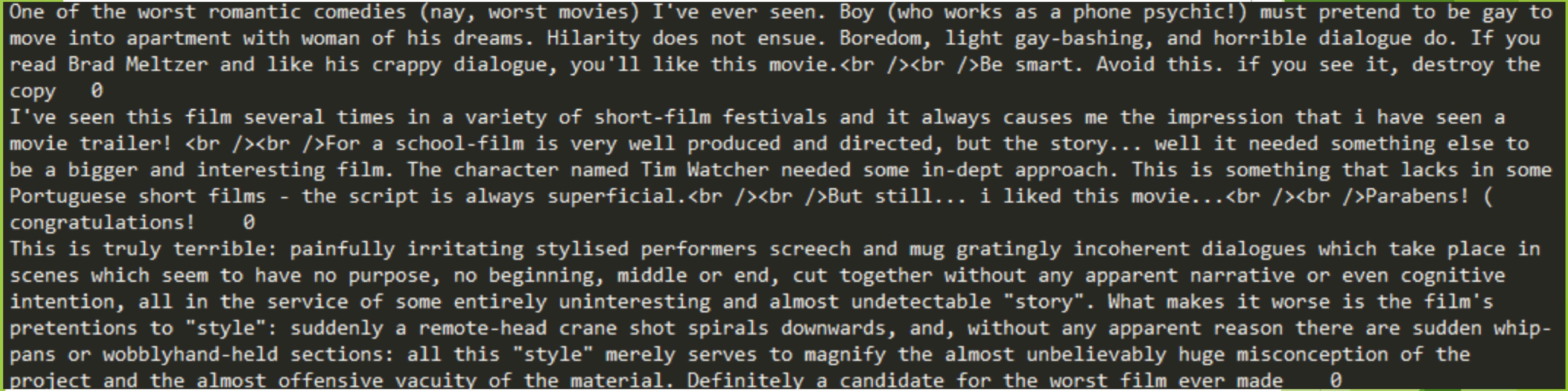
Dataset – IMDB movie reviews

- This dataset contains total of 50,000 movie reviews
- Each reviews is labelled as 0 or 1 : 1 means the review is positive and the viewer liked the movie and 0 means the review is negative and the viewer didn’t like the movie
- The sentiment to analyse to here can be termed as happy/unhappy or satisfied/unsatisfied or liked/disliked
- Training set – 15,000 reviews is segregated for training the classification model
- Validation set contains 10,000 reviews
- Test set contains 25,000 reviews



```
Yes, this gets the full ten stars. It's plain as day that this fill is genius. The universe sent Trent Harris a young, wonderfully
strange man one day and Harris caught him on tape, in all that true misfit glory that you just can't fake. Too bad it ended in
tragedy for the young man, if only an alternate ending could be written for that fellow's story. The other two steps in the trilogy
do retell the story, with Sean Penn and Crispin Glover in the roles of the young men, respectively. The world is expanded upon and
the strangeness is contextualized by the retelling, giving us a broader glimpse into growing up weird in vanilla America.
Recommended for anyone and everyone  1
Hello. This movie is.......well.......okay. Just kidding! ITS AWESOME! It's NOT a Block Buster smash hit. It's not meant to be. But
its a big hit in my world. And my sisters. We are rockin' Rollers. GO RAMONES!!!! This is a great movie............ For ME    1
This is a film that was very well done. I had heard mixed reviews while it was in production and have been waiting for its release!
Cheers to the director and all the actors. The supporting cast gave Eva Mendez what she needed to take this to the top. As everyone
else here states, the latter portion of the film is riveting. Katie Cassidy did an amazing job with her character, being she had not
done a lot of work when this film was made. She has quite the career ahead of her. I was amazed at her performance. I completely
enjoyed the film, questioned my values in life and priorities, and am a better person for it! A great message lies within the film.
Release it so all can enjoy   1
```

Sample from dataset of a positive review



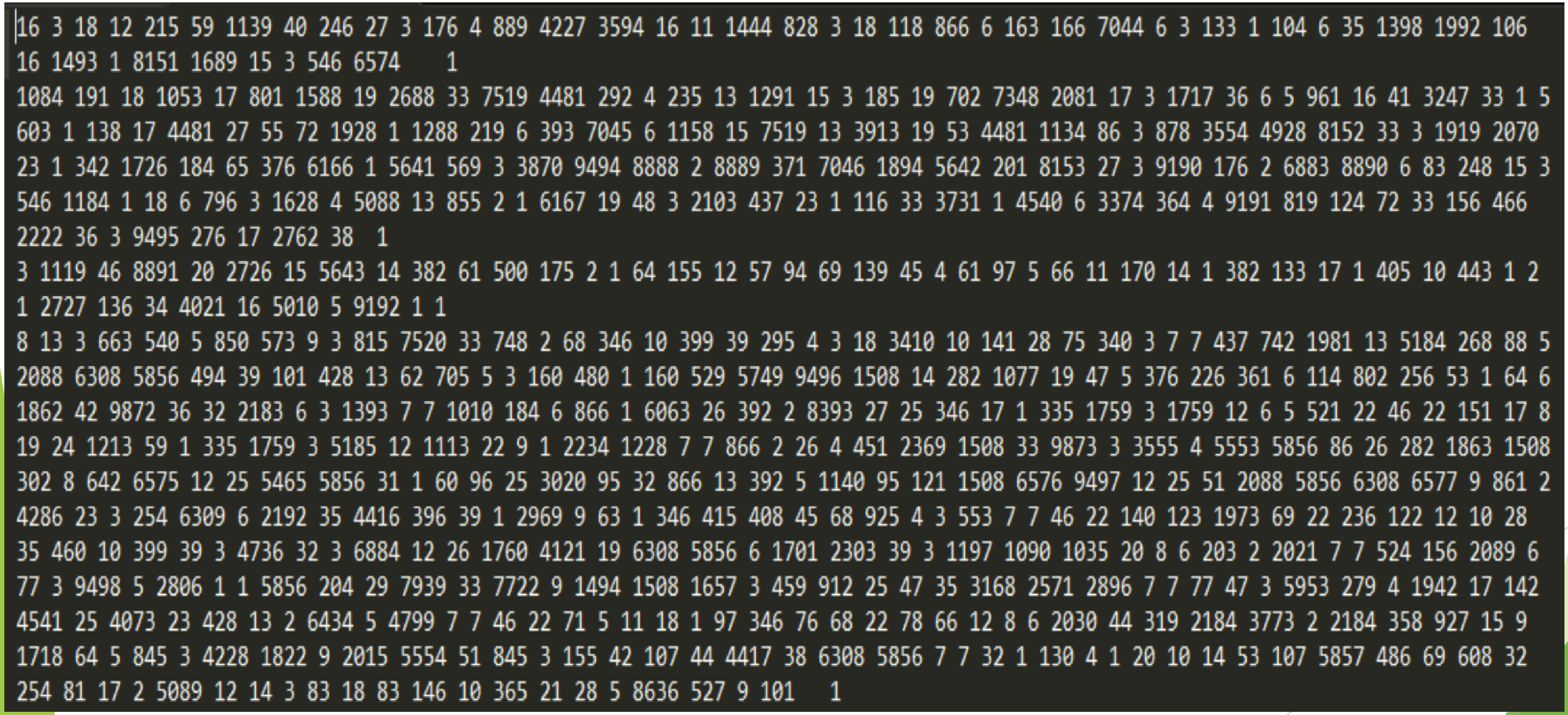
```
One of the worst romantic comedies (nay, worst movies) I've ever seen. Boy (who works as a phone psychic!) must pretend to be gay to
move into apartment with woman of his dreams. Hilarity does not ensue. Boredom, light gay-bashing, and horrible dialogue do. If you
read Brad Meltzer and like his crappy dialogue, you'll like this movie.<br /><br />Be smart. Avoid this. if you see it, destroy the
copy	0
I've seen this film several times in a variety of short-film festivals and it always causes me the impression that i have seen a
movie trailer! <br /><br />For a school-film is very well produced and directed, but the story... well it needed something else to
be a bigger and interesting film. The character named Tim Watcher needed some in-dept approach. This is something that lacks in some
Portuguese short films - the script is always superficial.<br /><br />But still... i liked this movie...<br /><br />Parabens! (
congratulations!	0
This is truly terrible: painfully irritating stylised performers screech and mug gratingly incoherent dialogues which take place in
scenes which seem to have no purpose, no beginning, middle or end, cut together without any apparent narrative or even cognitive
intention, all in the service of some entirely uninteresting and almost undetectable "story". What makes it worse is the film's
pretentions to "style": suddenly a remote-head crane shot spirals downwards, and, without any apparent reason there are sudden whip-
pans or wobblyhand-held sections: all this "style" merely serves to magnify the almost unbelievably huge misconception of the
project and the almost offensive vacuity of the material. Definitely a candidate for the worst film ever made	0
```

Sample from dataset of a negative review

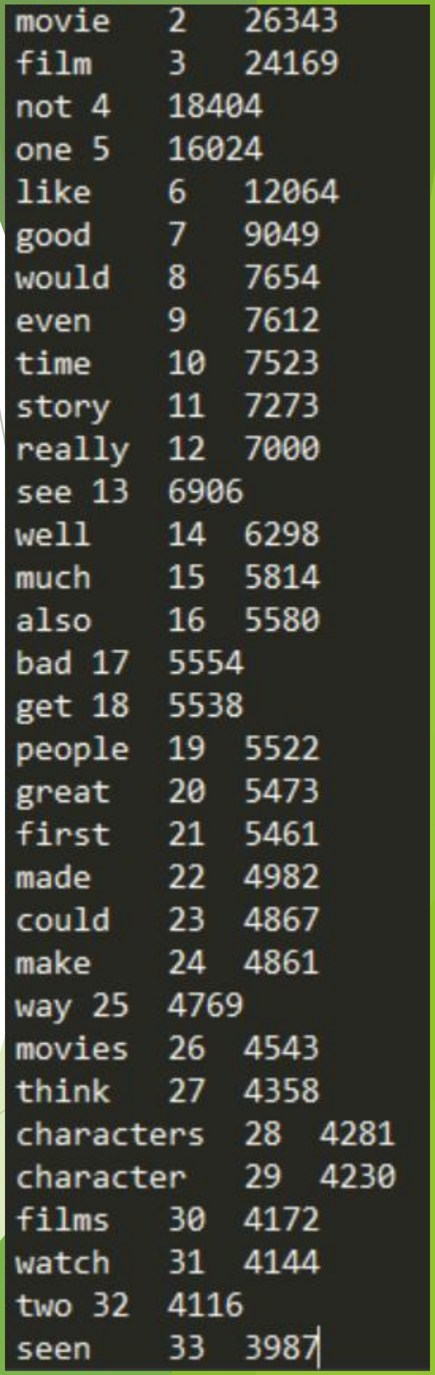
Dataset pre-processing

- Removal of everything else except words (numbers, special characters etc.)
- Conversion into lower case
- A dictionary is created from the words of raw training dataset
- In the dictionary, each unique word is assigned an ID and alongside the frequency of occurrence of that word is also stored
- Vocabulary is constructed from the top 10000 most occurring words in the training dataset
- Finally, the dataset is created by replacing each word in the raw dataset with its ID and ignoring those words which doesn’t exist in the vocabulary

Sample of Vocabulary and processed dataset



```
16 3 18 12 215 59 1139 40 246 27 3 176 4 889 4227 3594 16 11 1444 828 3 18 118 866 6 163 166 7044 6 3 133 1 104 6 35 1398 1992 106
16 1493 1 8151 1689 15 3 546 6574	1
1084 191 18 1053 17 801 1588 19 2688 33 7519 4481 292 4 235 13 1291 15 3 185 19 702 7348 2081 17 3 1717 36 6 5 961 16 41 3247 33 1 5
603 1 138 17 4481 27 55 72 1928 1 1288 219 6 393 7045 6 1158 15 7519 13 3913 19 53 4481 1134 86 3 878 3554 4928 8152 33 3 1919 2070
23 1 342 1726 184 65 376 6166 1 5641 569 3 3870 9494 8888 2 8889 371 7046 1894 5642 201 8153 27 3 9190 176 2 6883 8890 6 83 248 15 3
546 1184 1 18 6 796 3 1628 4 5088 13 855 2 1 6167 19 48 3 2103 437 23 1 116 33 3731 1 4540 6 3374 364 4 9191 819 124 72 33 156 466
2222 36 3 9495 276 17 2762 38	1
3 1119 46 8891 20 2726 15 5643 14 382 61 500 175 2 1 64 155 12 57 94 69 139 45 4 61 97 5 66 11 170 14 1 382 133 17 1 405 10 443 1 2
1 2727 136 34 4021 16 5010 5 9192 1 1
8 13 3 663 540 5 850 573 9 3 815 7520 33 748 2 68 346 10 399 39 295 4 3 18 3410 10 141 28 75 340 3 7 7 437 742 1981 13 5184 268 88 5
2088 6308 5856 494 39 101 428 13 62 705 5 3 160 480 1 160 529 5749 9496 1508 14 282 1077 19 47 5 376 226 361 6 114 802 256 53 1 64 6
1862 42 9872 36 32 2183 6 3 1393 7 7 1010 184 6 866 1 6063 26 392 2 8393 27 25 346 17 1 335 1759 3 1759 12 6 5 521 22 46 22 151 17 8
19 24 1213 59 1 335 1759 3 5185 12 1113 22 9 1 2234 1228 7 7 866 2 26 4 451 2369 1508 33 9873 3 3555 4 5553 5856 86 26 282 1863 1508
302 8 642 6575 12 25 5465 5856 31 1 60 96 25 3020 95 32 866 13 392 5 1140 95 121 1508 6576 9497 12 25 51 2088 5856 6308 6577 9 861 2
4286 23 3 254 6309 6 2192 35 4416 396 39 1 2969 9 63 1 346 415 408 45 68 925 4 3 553 7 7 46 22 140 123 1973 69 22 236 122 12 10 28
35 460 10 399 39 3 4736 32 3 6884 12 26 1760 4121 19 6308 5856 6 1701 2303 39 3 1197 1090 1035 20 8 6 203 2 2021 7 7 524 156 2089 6
77 3 9498 5 2806 1 1 5856 204 29 7939 33 7722 9 1494 1508 1657 3 459 912 25 47 35 3168 2571 2896 7 7 77 47 3 5953 279 4 1942 17 142
4541 25 4073 23 428 13 2 6434 5 4799 7 7 46 22 71 5 11 18 1 97 346 76 68 22 78 66 12 8 6 2030 44 319 2184 3773 2 2184 358 927 15 9
1718 64 5 845 3 4228 1822 9 2015 5554 51 845 3 155 42 107 44 4417 38 6308 5856 7 7 32 1 130 4 1 20 10 14 53 107 5857 486 69 608 32
254 81 17 2 5089 12 14 3 83 18 83 146 10 365 21 28 5 8636 527 9 101	1
```



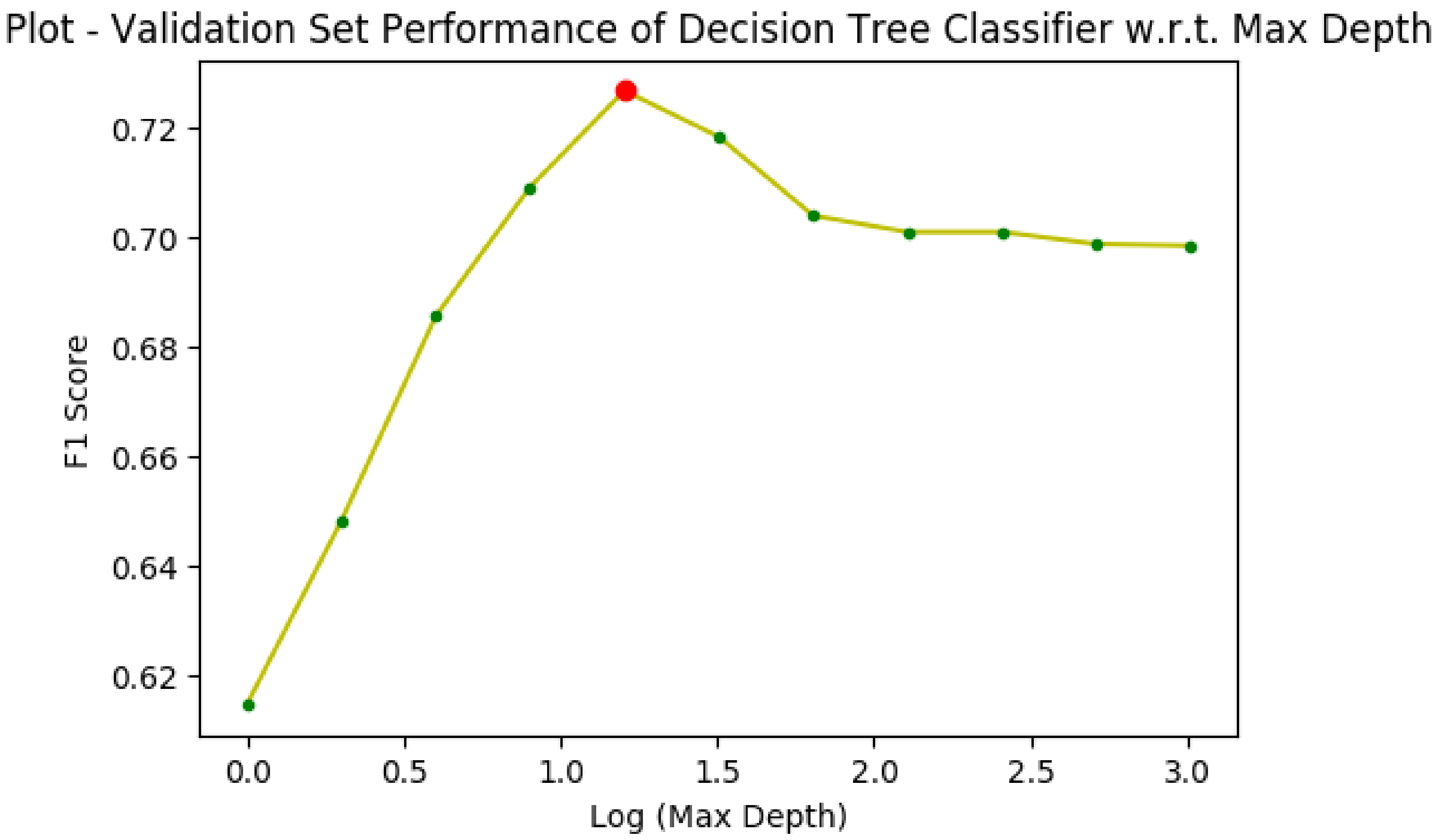
```
movie	2	26343
film	3	24169
not 4	18404
one 5	16024
like	6	12064
good	7	9049
would	8	7654
even	9	7612
time	10	7523
story	11	7273
really	12	7000
see 13	6906
well	14	6298
much	15	5814
also	16	5580
bad 17	5554
get 18	5538
people	19	5522
great	20	5473
first	21	5461
made	22	4982
could	23	4867
make	24	4861
way 25	4769
movies	26	4543
think	27	4358
characters	28	4281
character	29	4230
films	30	4172
watch	31	4144
two 32	4116
seen	33	3987
```

Classification model training:

Binary bag of words

Decision Tree Classifier

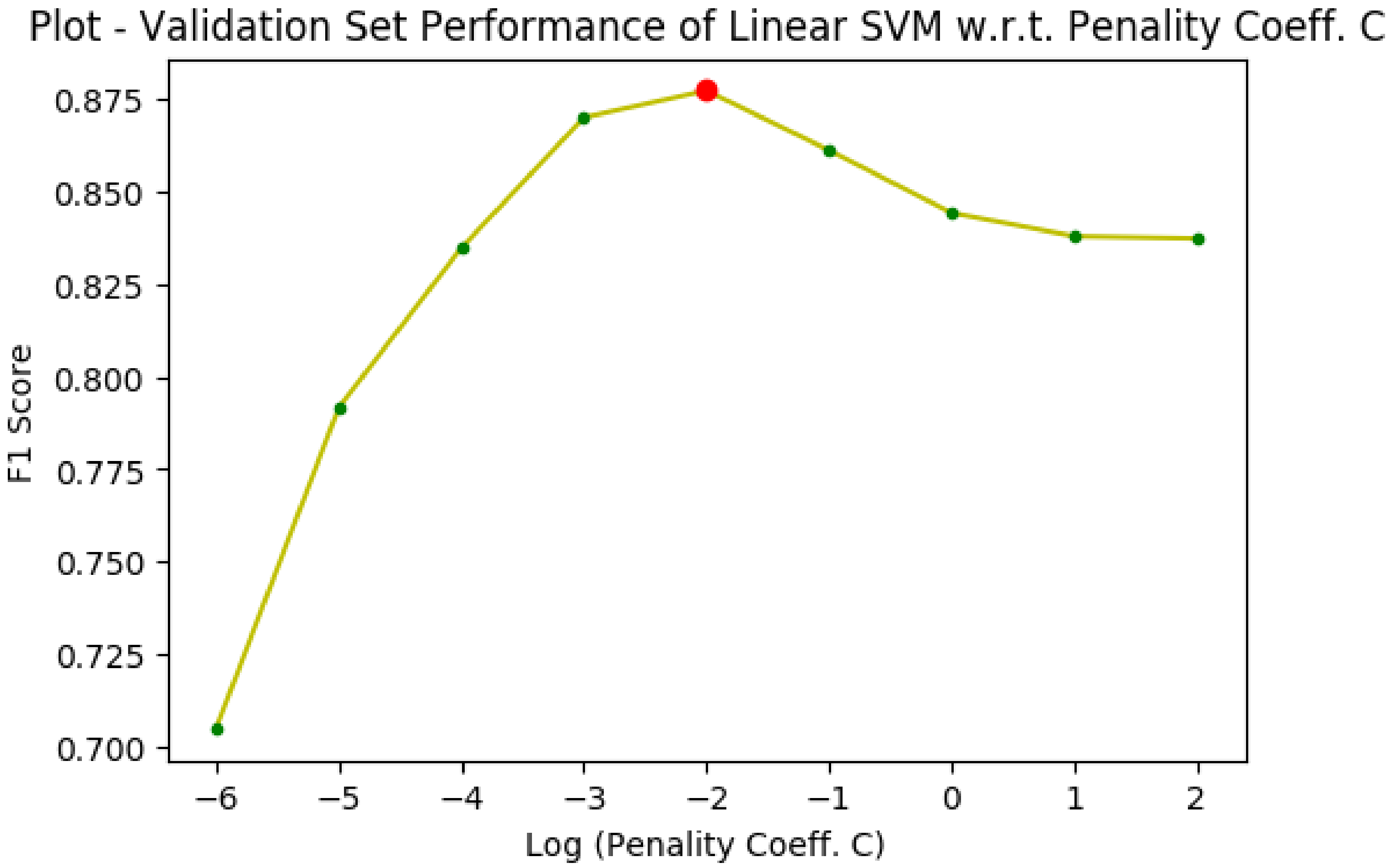
- Max_depth with best performance : 16
- Training set F1 score : 0.8338
- Validation set F1 score : 0.7261
- Test set F1 score : 0.731



Plot - Validation Set Performance of Decision Tree Classifier w.r.t. Max Depth

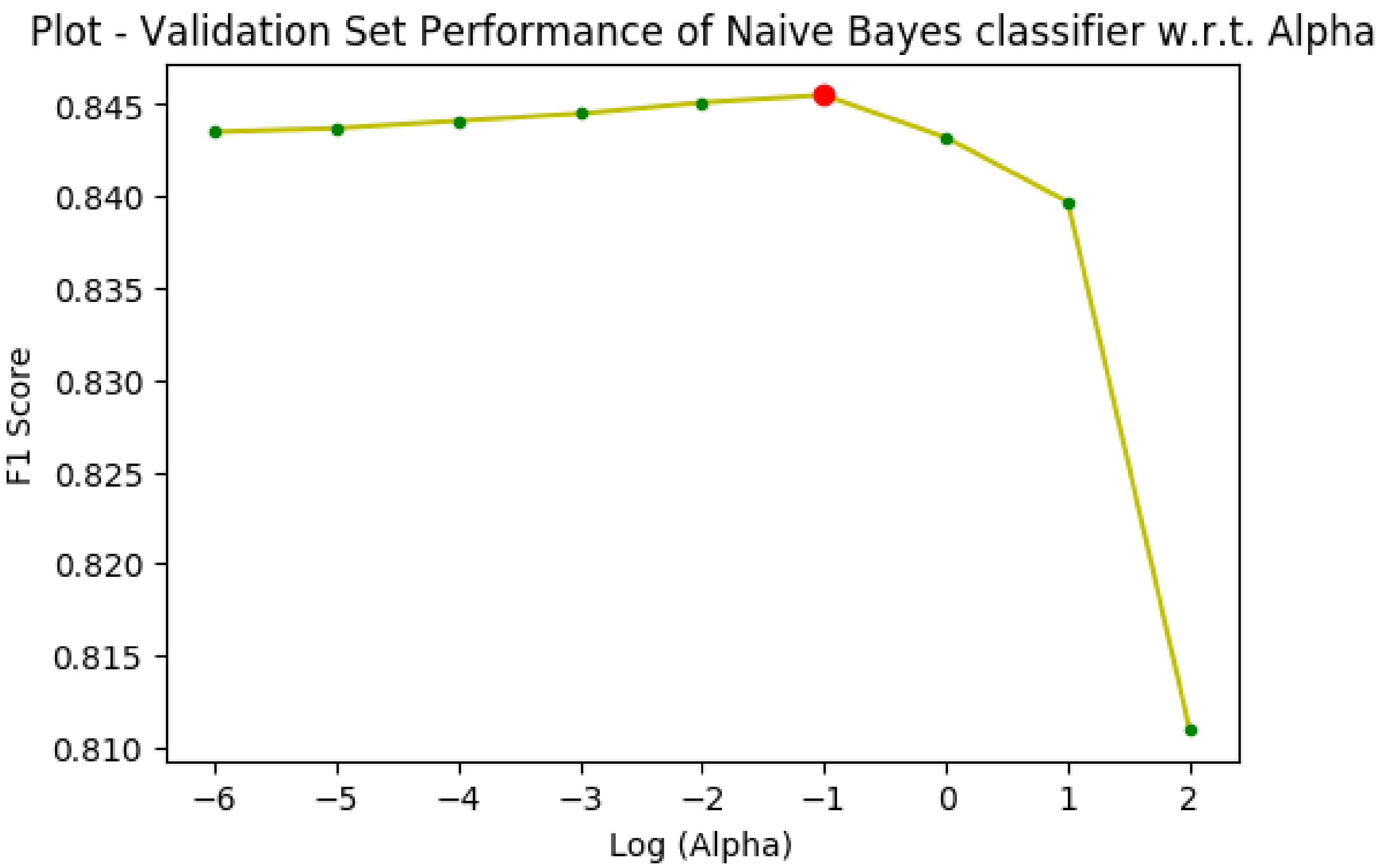
Linear Support Vector Machine

- Penalty coeff with best performance : 0.01
- Training set F1 score : 0.9639
- Validation set F1 score : 0.8772
- Test set F1 score : 0.87032



Bernoulli Naïve Bayes Classifier

- Alpha with best performance : 0.1
- Training set F1 score : 0.8719
- Validation set F1 score : 0.8455
- Test set F1 score : 0.8326

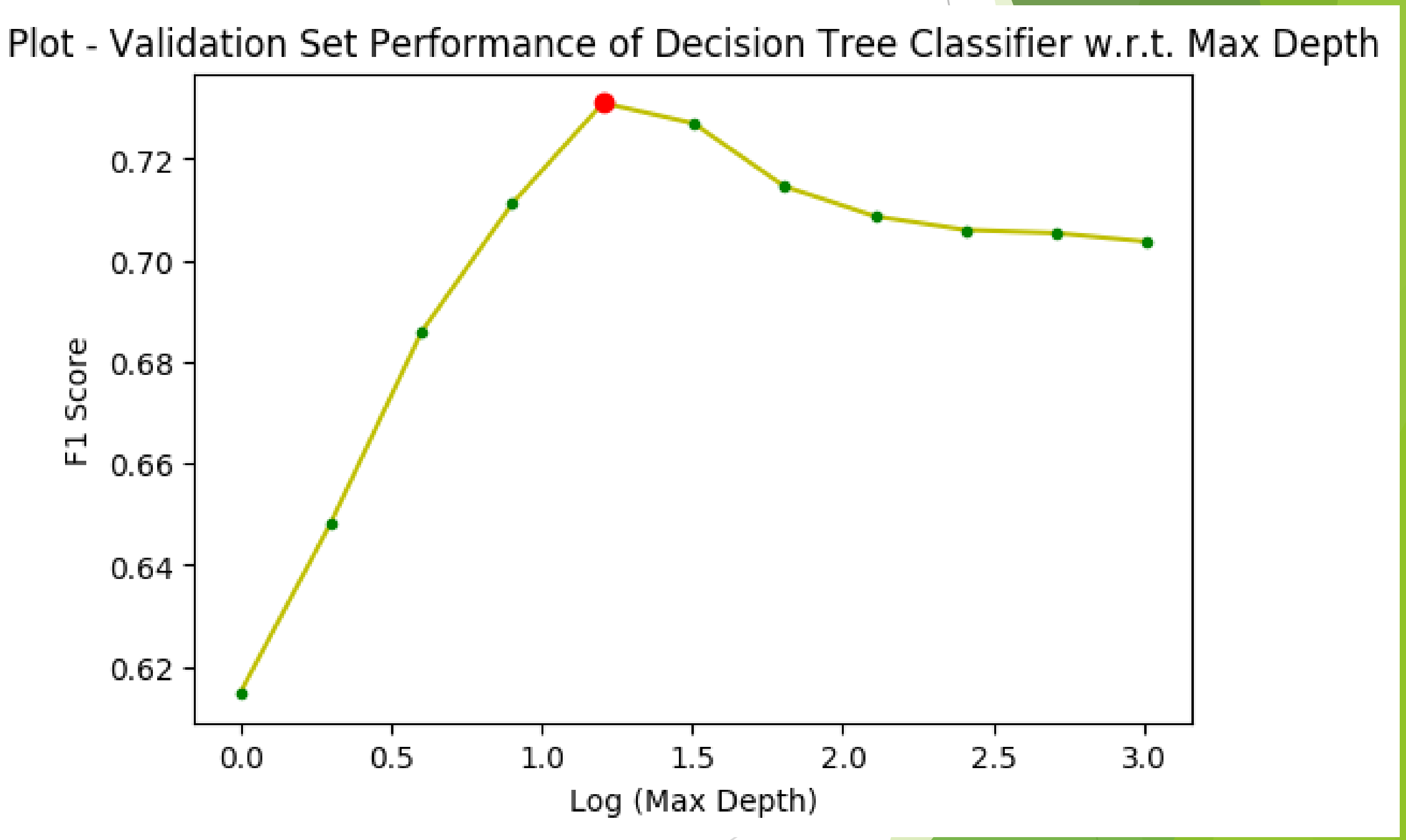


Refined Dataset

Preprocessed the dataset further by removing the stop words and single letters

Decision Tree Classifier

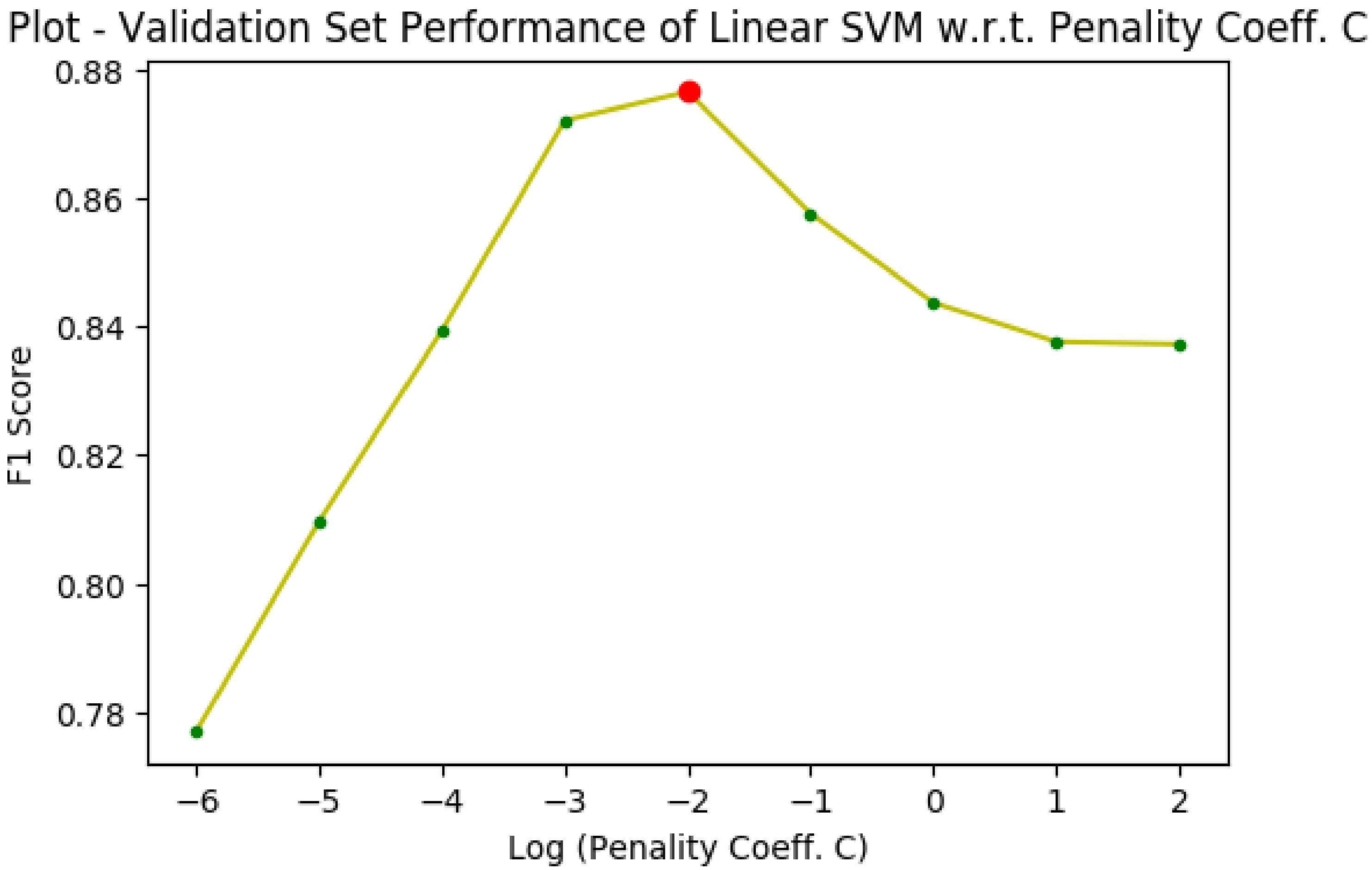
- Max_depth with best performance : 16
- Training set F1 score : 0.8274
- Validation set F1 score : 0.73
- Test set F1 score : 0.73224



Plot - Validation Set Performance of Decision Tree Classifier w.r.t. Max Depth

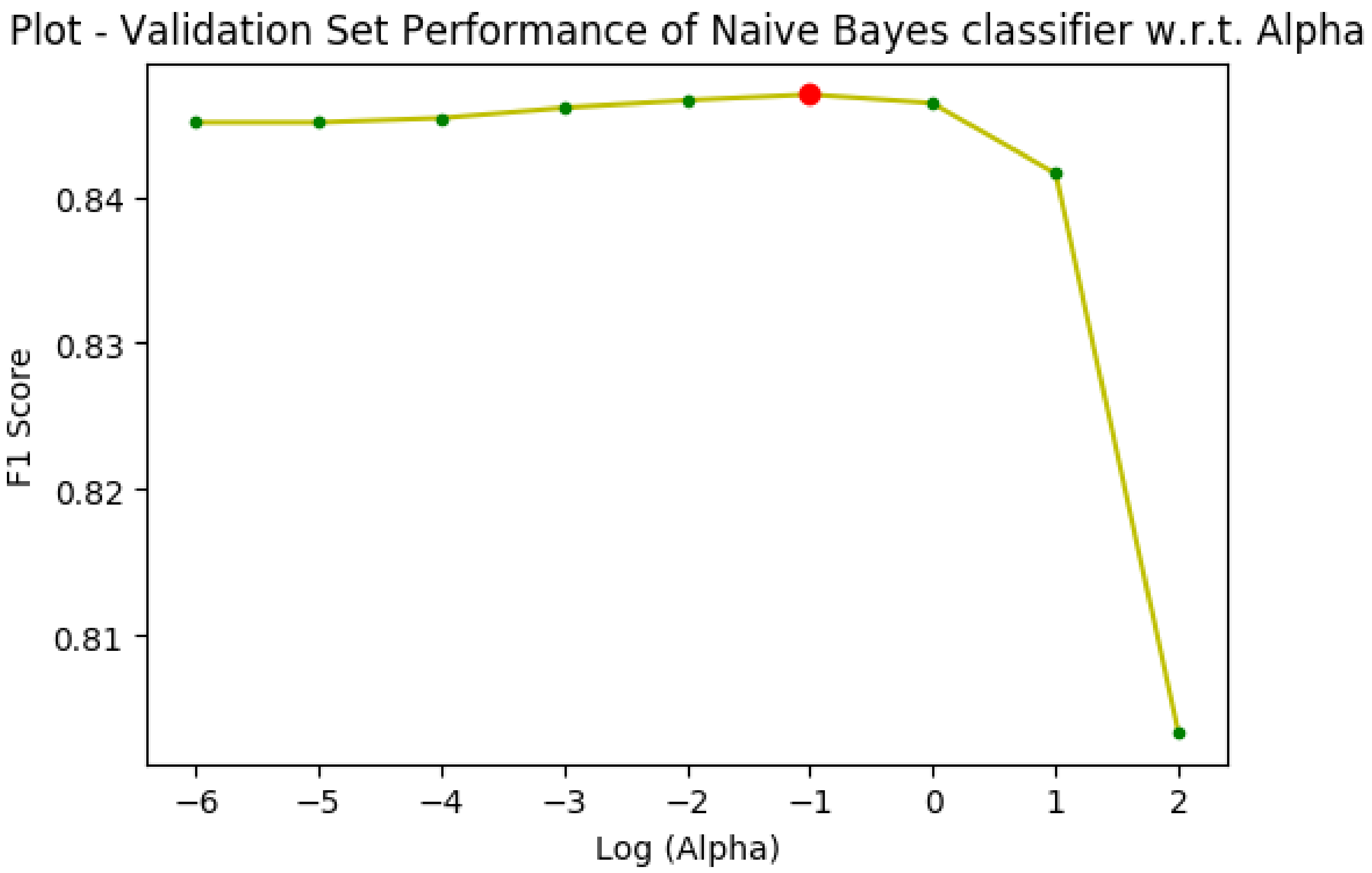
Linear Support Vector Machine

- Penalty coeff with best performance : 0.01
- Training set F1 score : 0.9633
- Validation set F1 score : 0.8765
- Test set F1 score : 0.87



Bernoulli Naïve Bayes Classifier

- Alpha with best performance : 0.1
- Training set F1 score : 0.8725
- Validation set F1 score : 0.847
- Test set F1 score : 0.83

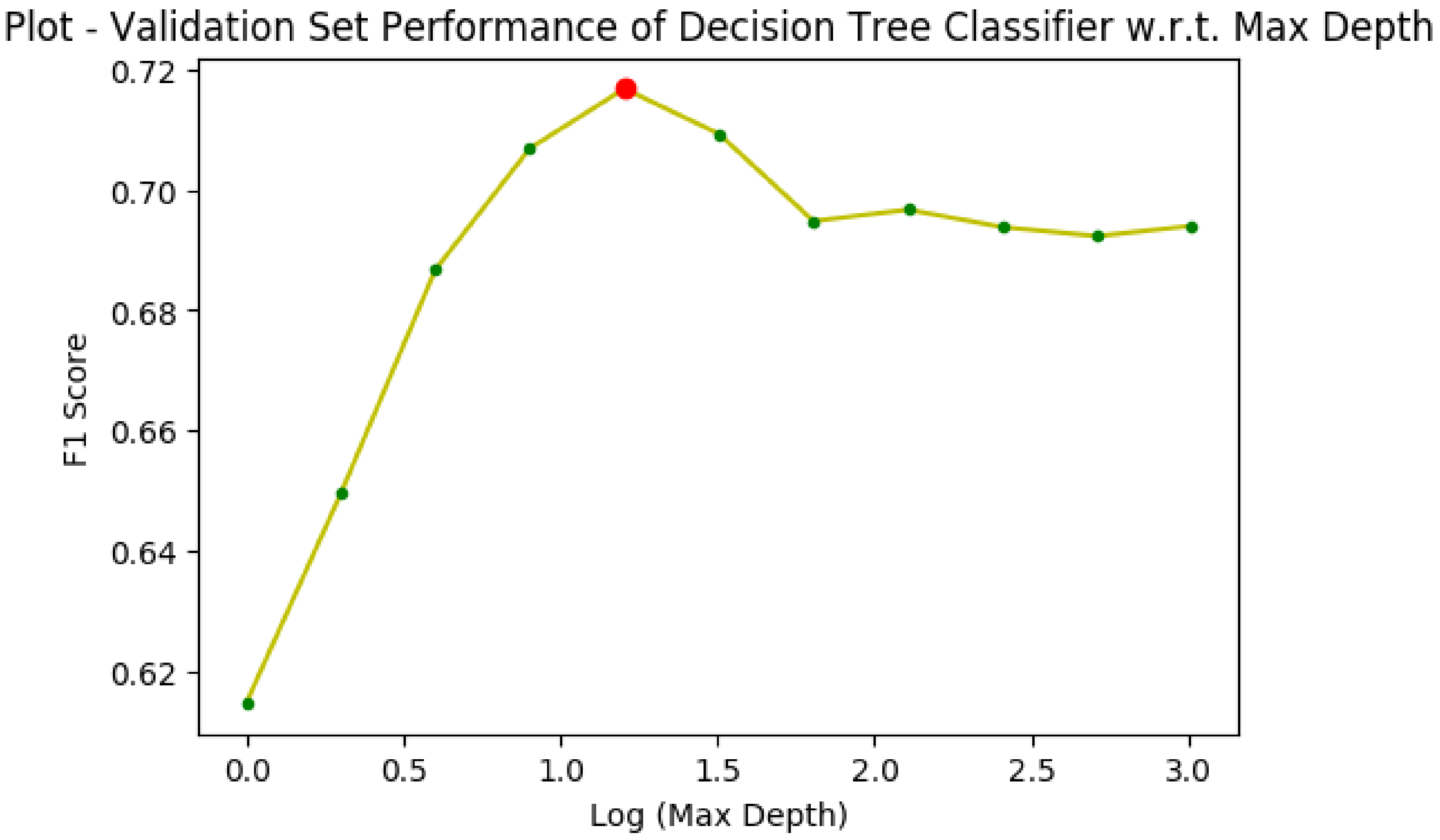


Classification model training:

Frequency bag of words

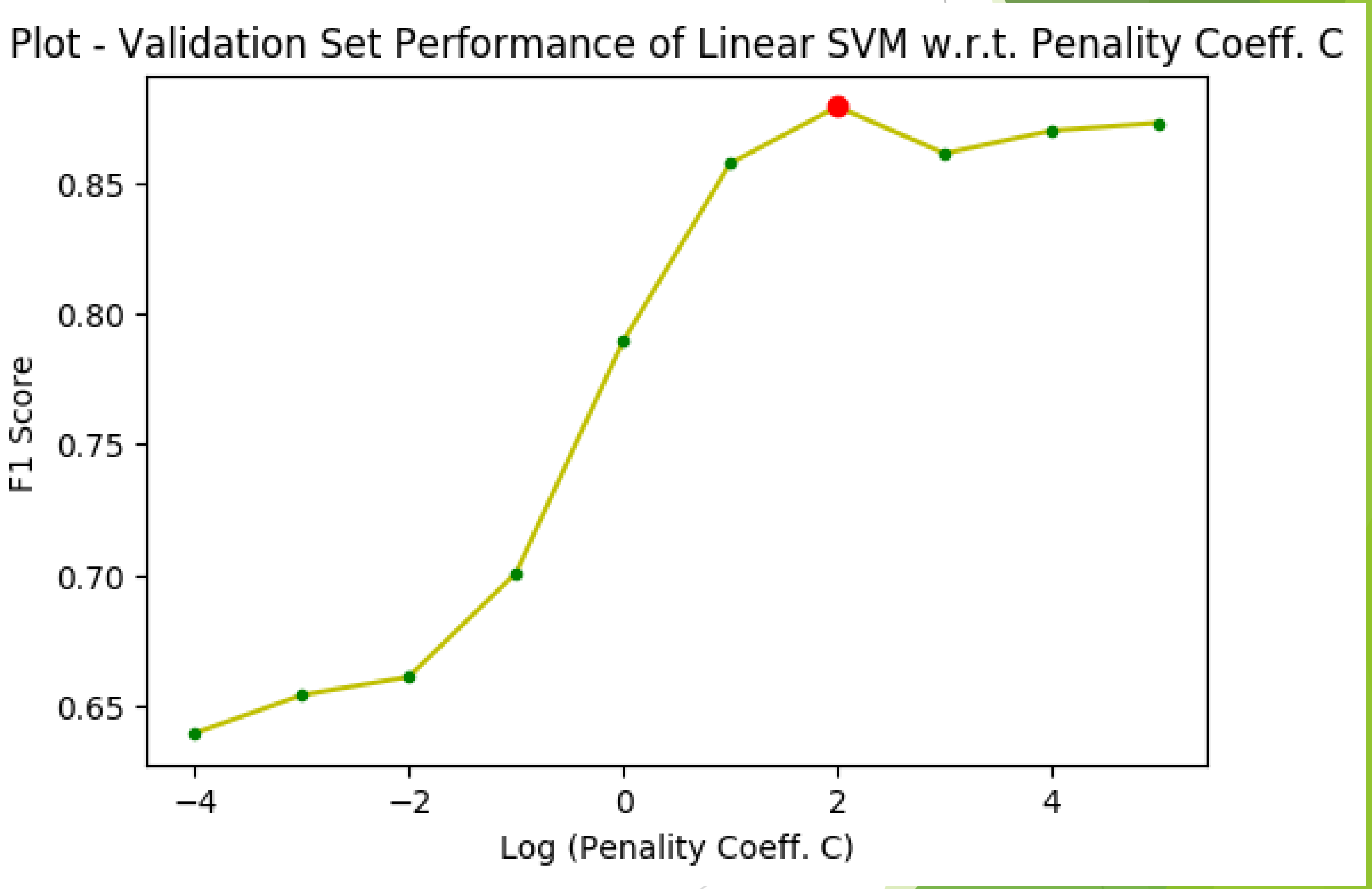
Decision Tree Classifier

- Max_depth with best performance : 16
- Training set F1 score : 0.8533
- Validation set F1 score : 0.7157
- Test set F1 score : 0.72452


Plot - Validation Set Performance of Decision Tree Classifier w.r.t. Max Depth

Linear Support Vector Machine

- Penalty coeff with best performance : 100.0
- Training set F1 score : 0.946933
- Validation set F1 score : 0.8794
- Test set F1 score : 0.87436



Gaussian Naïve Bayes Classifier

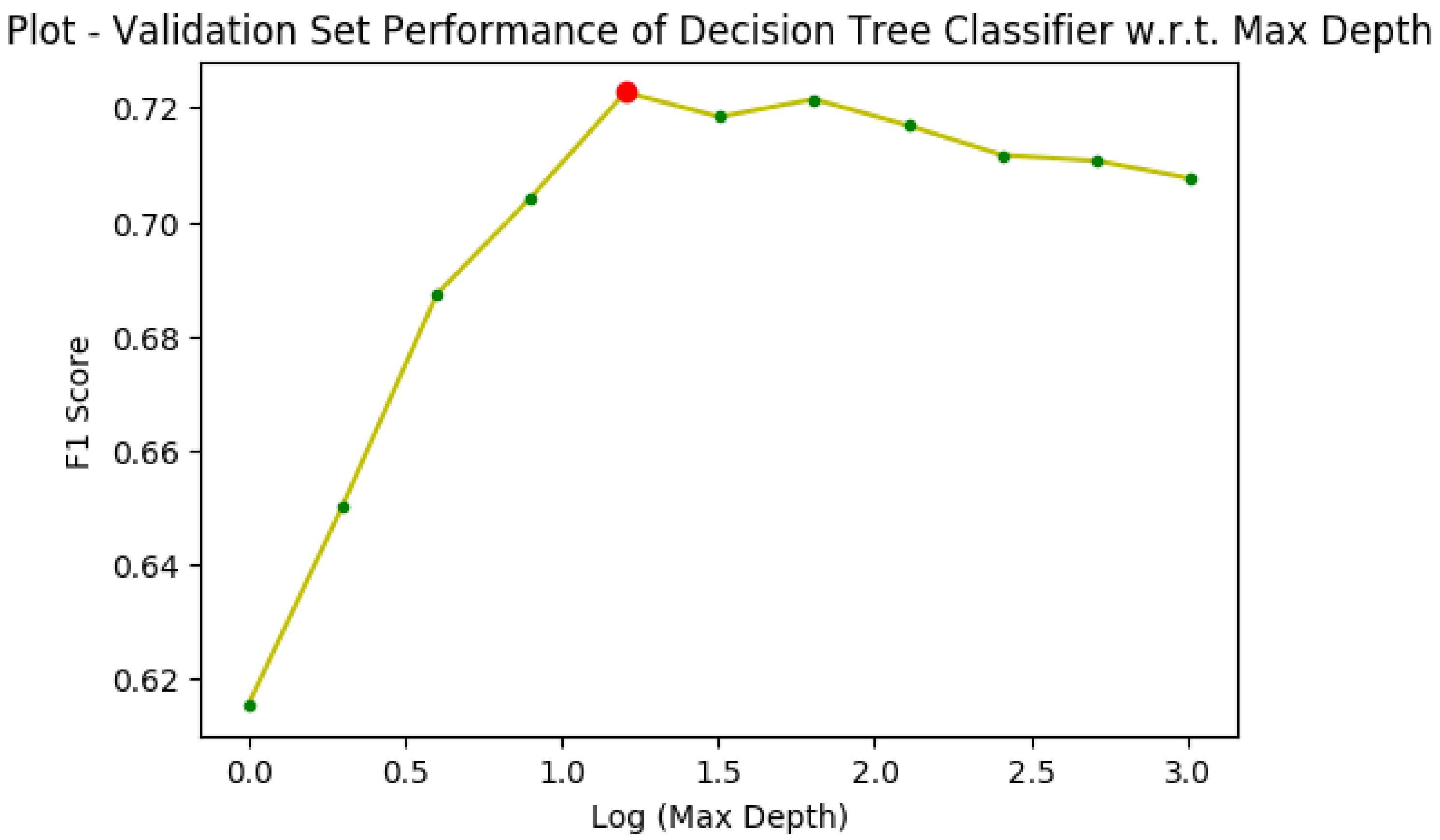
- Training set F1 score : 0.857066666667
- Validation set F1 score : 0.7521
- Test set F1 score : 0.68244

Refined Dataset

Preprocessed the data by removing the stop words and single letters

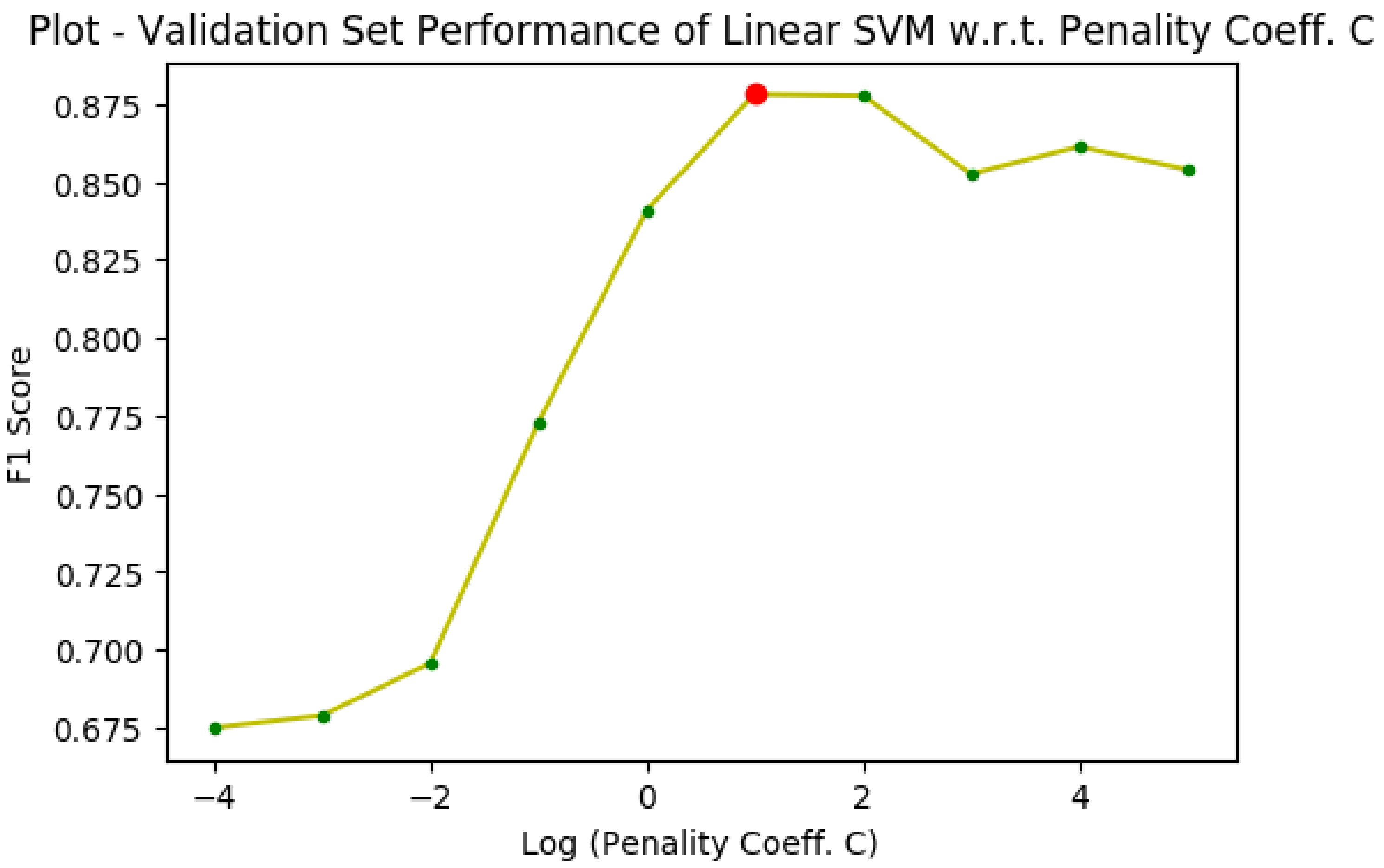
Decision Tree Classifier

- Max_depth with best performance : 16
- Training set F1 score : 0.8362
- Validation set F1 score : 0.723
- Test set F1 score : 0.7312



Linear Support Vector Machine

- Penalty coeff with best performance : 10.0
- Training set F1 score : 0.9223
- Validation set F1 score : 0.8782
- Test set F1 score : 0.8721



Gaussian Naïve Bayes Classifier

- Training set F1 score : 0.8641
- Validation set F1 score : 0.7538
- Test set F1 score : 0.6818