The Document Vectors Using Cosine Similarity Revisited

Name: Mohammad Haghighat

Course: Natural Language Processing

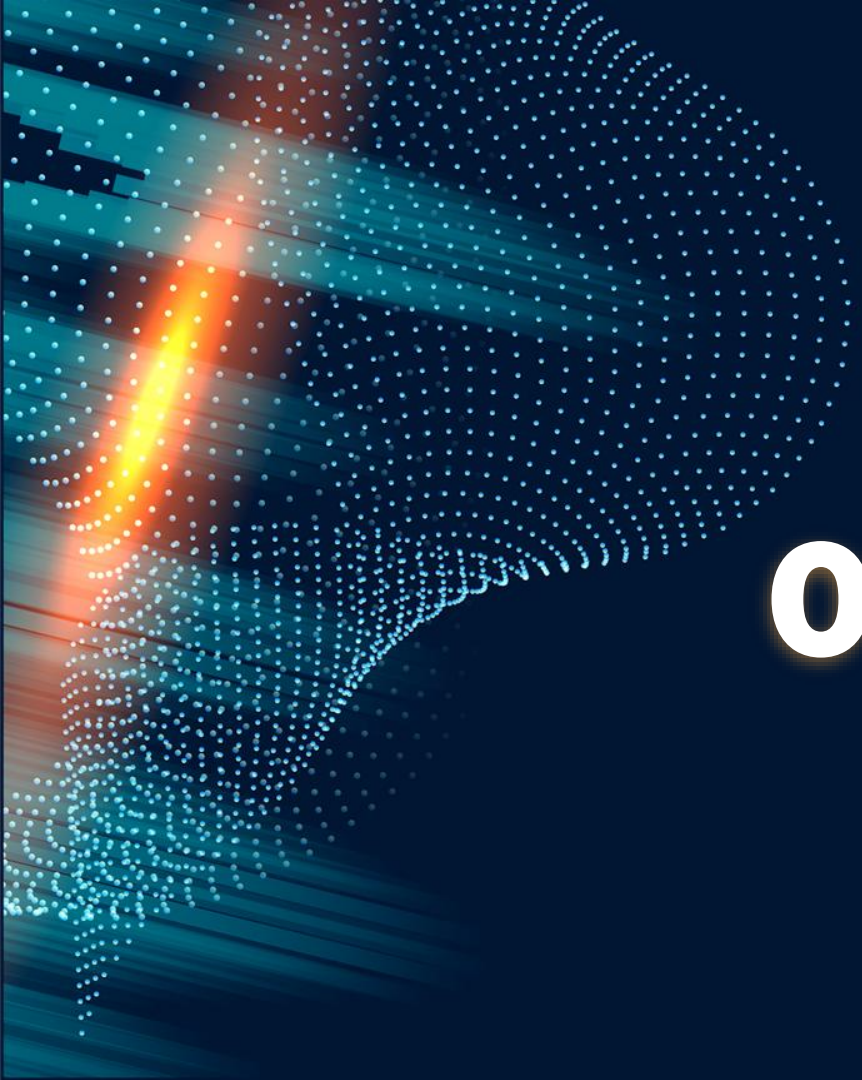
Teacher: Dr Behrouz Minaei

Teaching assistant: Mohammad Mousavi

03 Feb 2025

CONTENTS OF THIS PAPER

There's what you'll find in this paper:



01 Re-evaluation of the ensemble

Re-evaluation of the ensemble

Paper: Tan Thongtan and Tanasanee Phienthrakul. 2019

Accuracy: **97/42%**

State-of-the-art

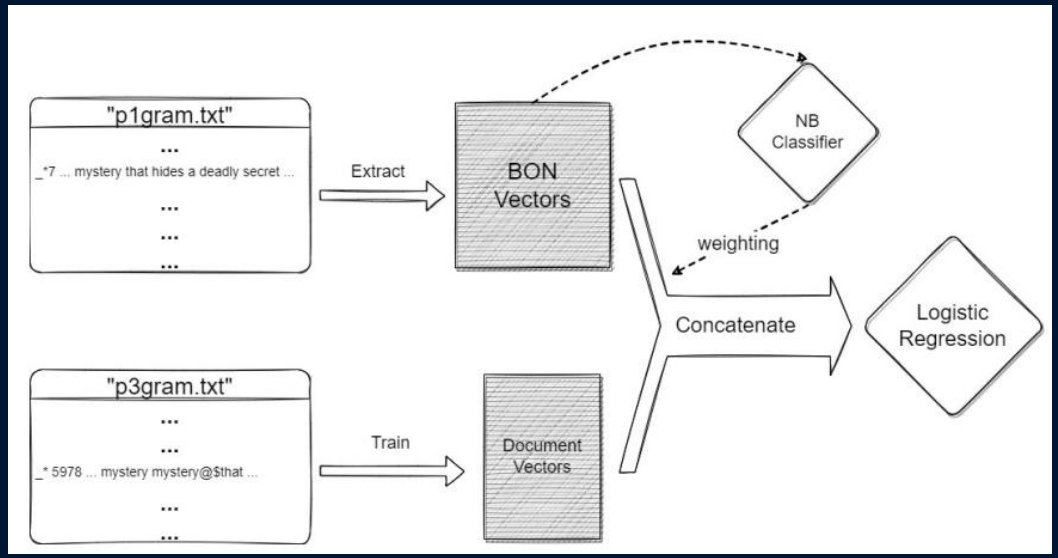
IMDB Dataset

Ensemble NBweighted BON and the DV-ngrams-cosine

Use Logistic Regression

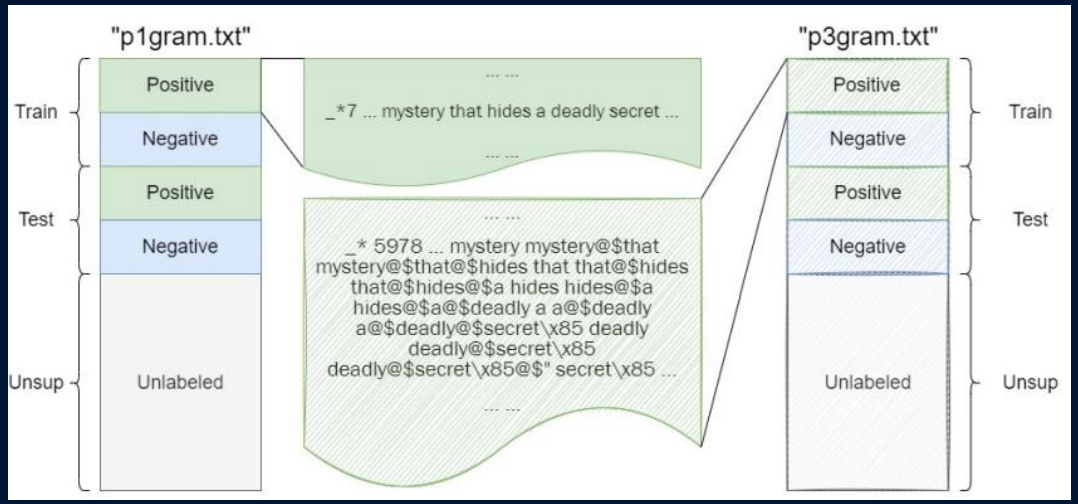
Find a bug in the evaluation procedure

Concatenate two vector representations (DV-ngrams-cosine and BON)

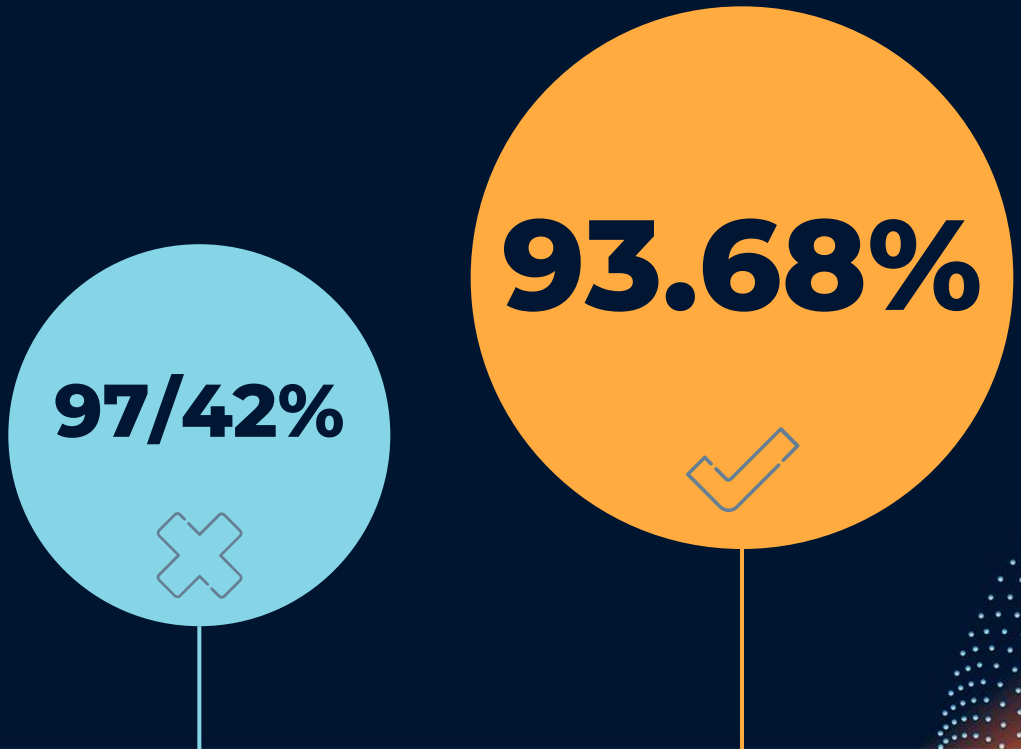


Find a bug in the evaluation procedure

Concatenate two vector representations (DV-ngrams-cosine and BON)

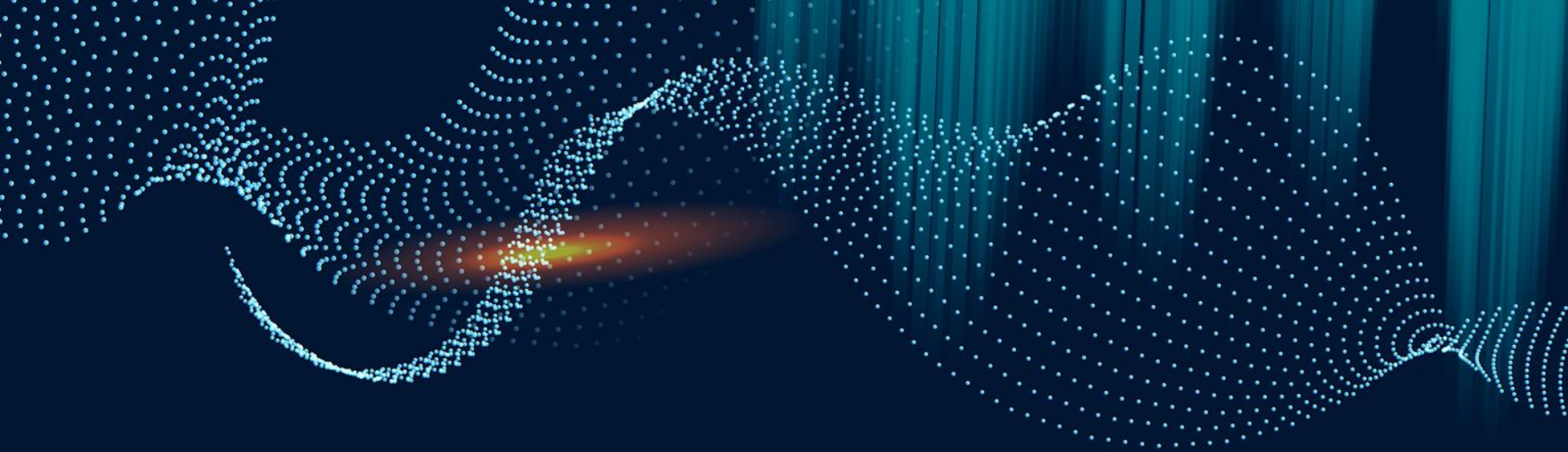


Real Accuracy



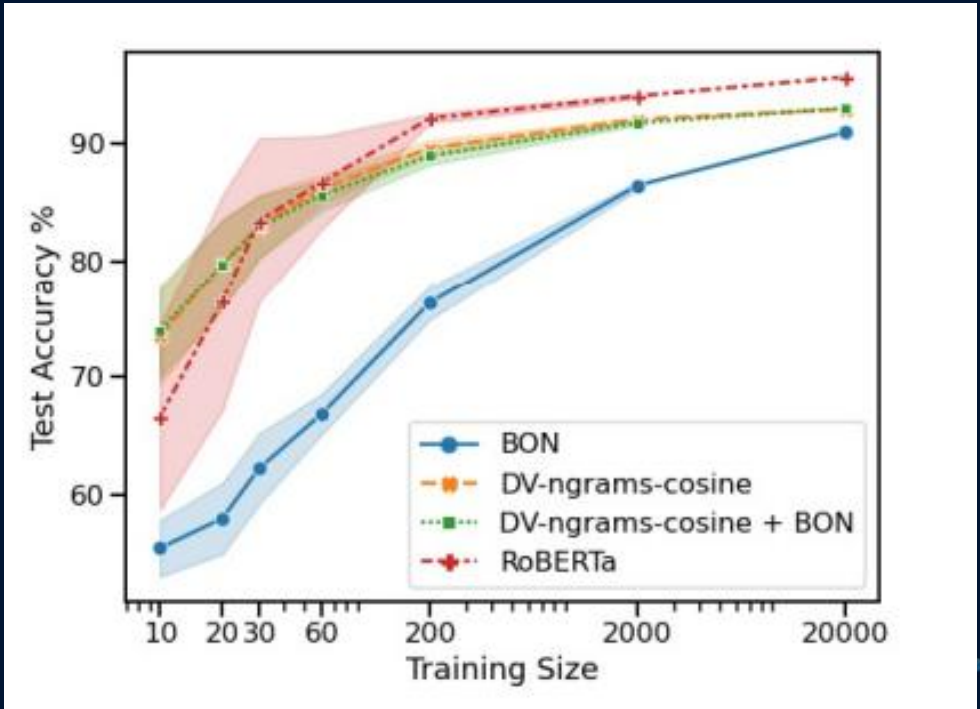
Real Accuracy

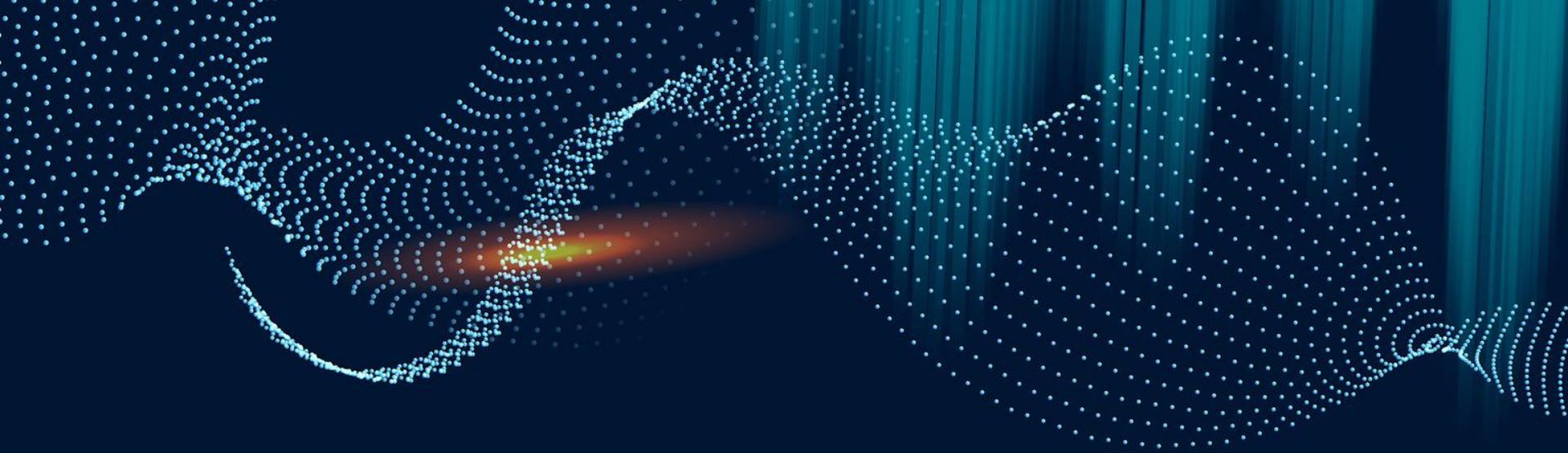
Model	Test Accuracy %
Models trained on the original training set of IMDB (25K)	
NB-weighted BON	91.29
DV-ngrams-cosine	93.13
DV-ngrams-cosine + NB-weighted BON (Thongtan and Phienthrakul, 2019)	#97.42
DV-ngrams-cosine + NB-weighted BON (re-evaluated)	93.68
Models trained using the train/dev split from (Suchin et al., 2020) (20K/5K)	
DV-ngrams-cosine with NB sub-sampling	93.36
RoBERTa	95.79
DV-ngrams-cosine + RoBERTa	95.92
DV-ngrams-cosine with NB sub-sampling + RoBERTa	95.94



02 Further analysis of performance

The performance of different models on training sets of different sizes





03 | NB Sub-Sampling

Naive Bayesian Sub-Sampling

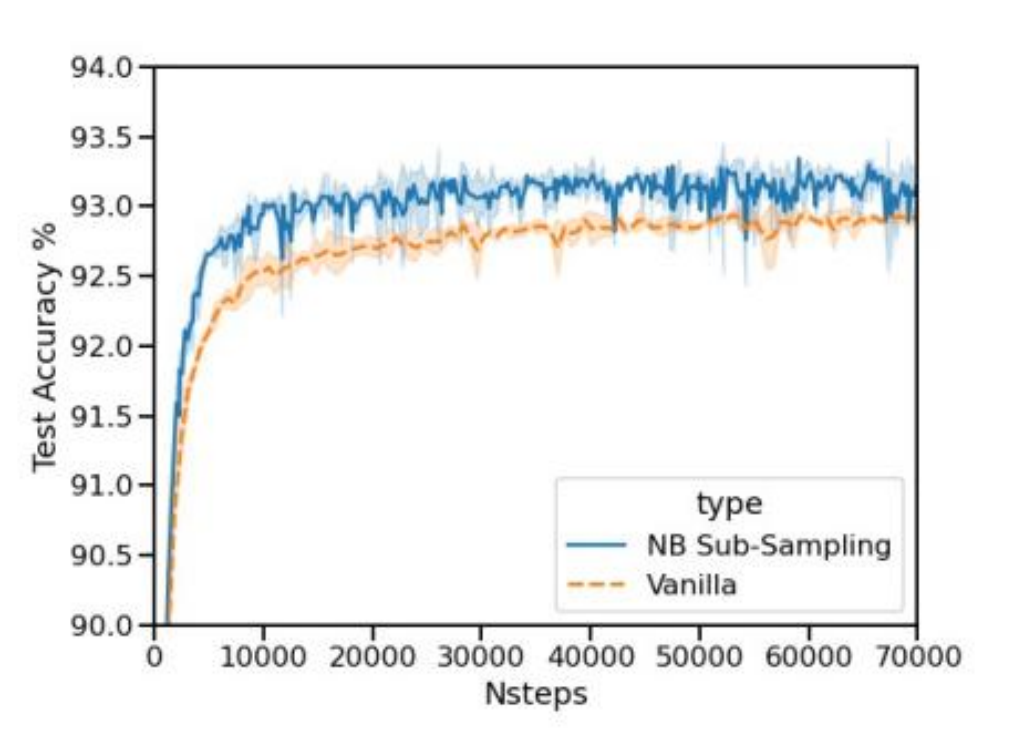
Naive Bayesian Classifier

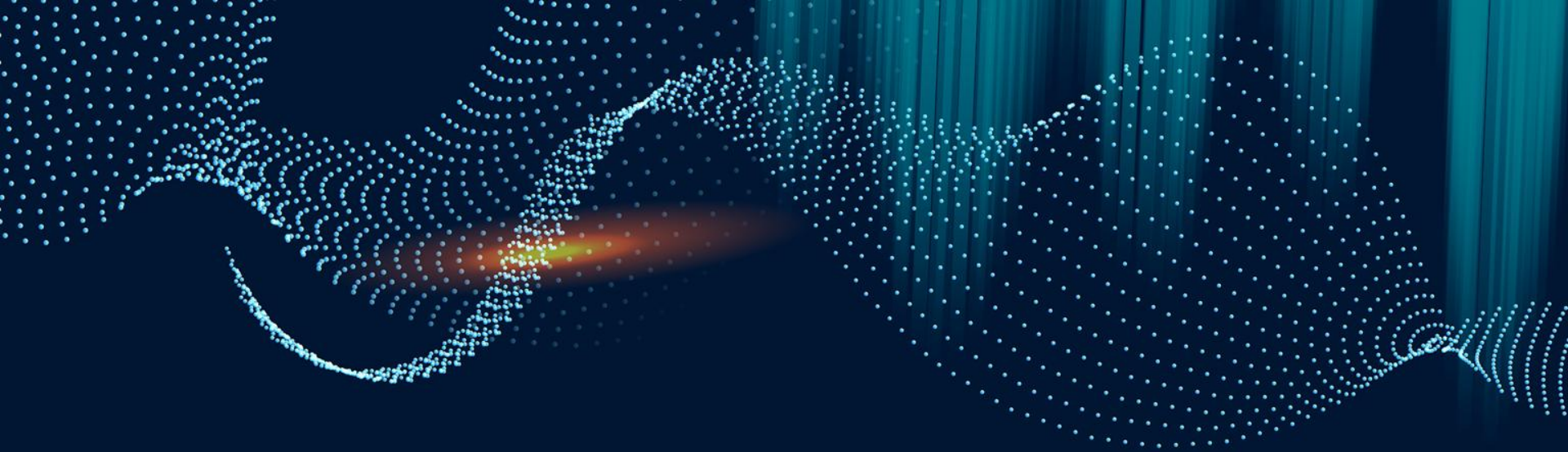
$$h_i = |\log p(f_i|y=1) - \log p(f_i|y=0)|$$

Sub-Sampling Probability

$$p(f_i) = min(\exp(h_i/n_a)/n_b, 1),$$

Naive Bayesian Sub-Sampling

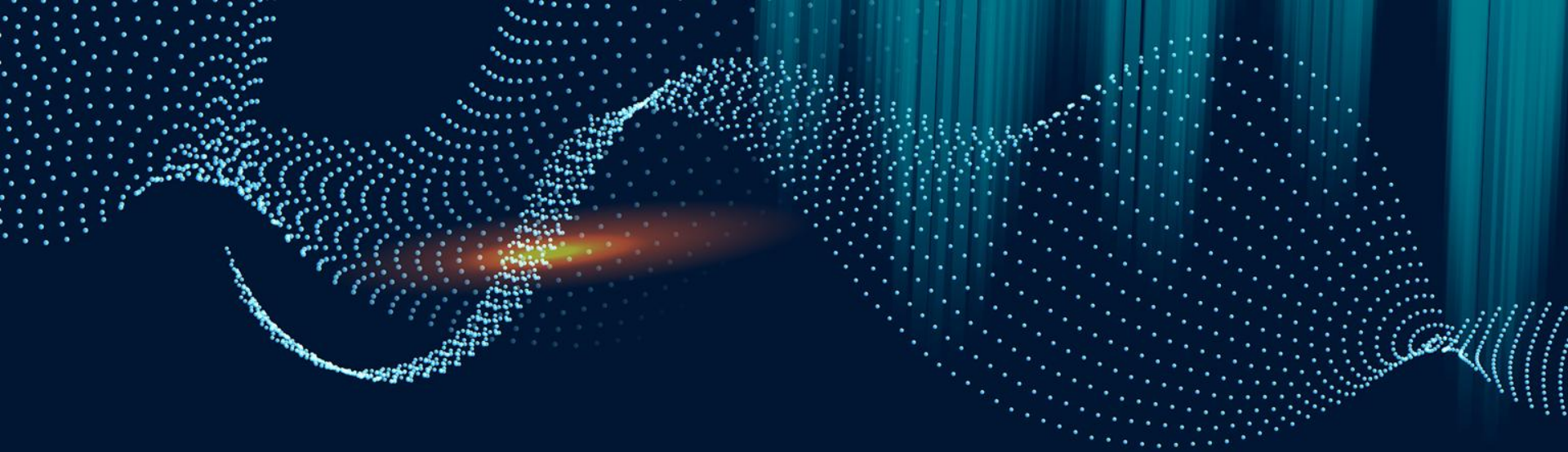




04 Ensemble DV-ngrams-cosine and RoBERTa

Ensemble DV-ngrams-cosine and RoBERTa

Model	Test Accuracy %
Models trained on the original training set of IMDB (25K)	
NB-weighted BON	91.29
DV-ngrams-cosine	93.13
DV-ngrams-cosine + NB-weighted BON (Thongtan and Phienthrakul, 2019)	#97.42
DV-ngrams-cosine + NB-weighted BON (re-evaluated)	93.68
Models trained using the train/dev split from (Suchin et al., 2020) (20K/5K)	
DV-ngrams-cosine with NB sub-sampling	93.36
RoBERTa	95.79
DV-ngrams-cosine + RoBERTa	95.92
DV-ngrams-cosine with NB sub-sampling + RoBERTa	95.94



05 My Experiments

Run

```
(base) PS C:\WINDOWS\system32> d:
(base) PS D:\> cd "D:\My Files\My Projects\Education\Uni\Master\NLP\Project\dv_cosine_revisited-main"
(base) PS D:\My Files\My Projects\Education\Uni\Master\NLP\Project\dv_cosine_revisited-main> python original_to_1gram.py
reading
checking
matched
(base) PS D:\My Files\My Projects\Education\Uni\Master\NLP\Project\dv_cosine_revisited-main> python test_with_origin.py
retrieving order
testing with the original order
testing with the correct order
testing with shuffled test set (within class)
100%| 30/30 [1:57:59<00:00, 235.99s/it]
testing with shuffled test set (whole)
 60%| 18/30 [1:08:29<20:02, 100.23s/it]
100%| 30/30 [2:13:26<00:00, 266.90s/it]
testing with shuffled train and test sets (inclass)
100%| 30/30 [44:08<00:00, 88.27s/it]
testing with shuffled train and test sets (whole)
100%| 30/30 [5:51:16<00:00, 702.55s/it]
saved report to test_logs\report.txt
finished
(base) PS D:\My Files\My Projects\Education\Uni\Master\NLP\Project\dv_cosine_revisited-main>
```

Results

Shuffling Scheme	Acc. Mean	Acc. Std.
original matching	97.42	
correct matching	93.68	
test set in-class (A)	96.58	0.07
test set cross-class (B)	61.80	0.25
train/test in-class (C)	97.43	0.08
train/test cross-class (D)	91.64	0.08

Table 2: Test accuracy for different shuffling schemes.

```
original score: 97.36
correct score: 93.58
shuffle test set in class: mean 96.73, std 0.08
shuffle whole test set: mean 64.17, std 0.23
shuffle train and test sets in class: mean 97.39, std 0.07
shuffle whole train and test sets: mean 91.60, std 0.06
```

Thanks