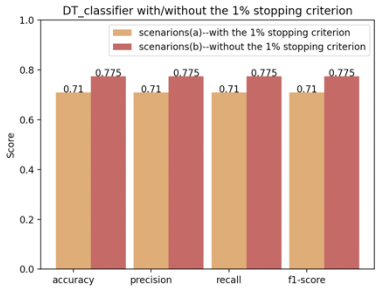
1.

(i)

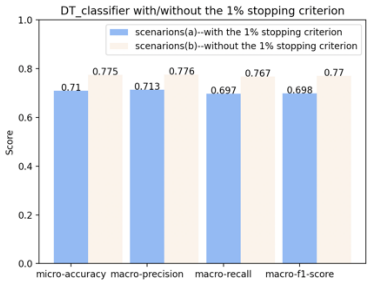
The results of data from *article.tsv*

Micro:



Micro	accuracy	precision	recall	F1
a	0.71	0.71	0.71	0.71
b	0.775	0.775	0.775	0.775

Macro:



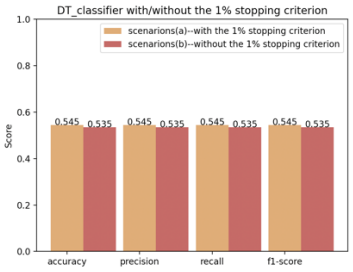
Macro	accuracy	precision	recall	F1
a	0.71	0.713	0.697	0.698
b	0.775	0.776	0.767	0.77

The DT without 1% stopping criterion have higher precision, recall and F1. In addition, their macro-precisions are all the highest in all figures.

(ii)

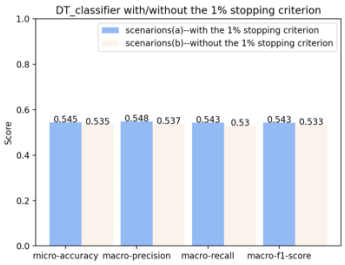
The results of data from *summaries.tsv*

Micro:



Micro	accuracy	precision	recall	F1
a	0.545	0.545	0.545	0.545
b	0.535	0.535	0.535	0.535

Macro:



Macro	accuracy	precision	recall	F1
a	0.545	0.548	0.543	0.543
b	0.535	0.537	0.53	0.533

The DT with 1% stopping criterion have higher precision, recall and F1. In addition, their macro-precisions are all the highest in all figures.

(iii)

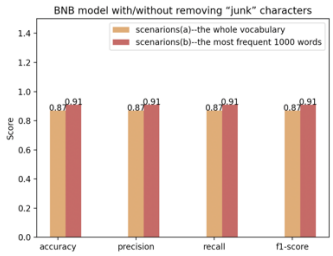
To compare the performance of DT model with 1% stopping criterion, we can find that using 1% stopping criterion to improve the accuracy or other figures is unstable.

2.

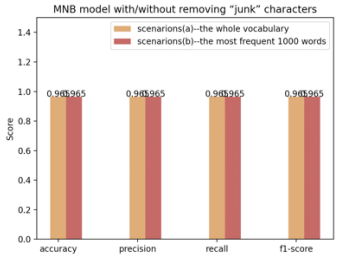
(i)

The results of data from *article.tsv*

Micro:

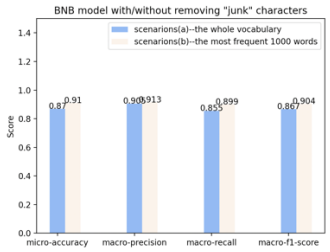


Micro	accuracy	precision	recall	F1
a	0.87	0.87	0.87	0.87
b	0.91	0.91	0.91	0.91

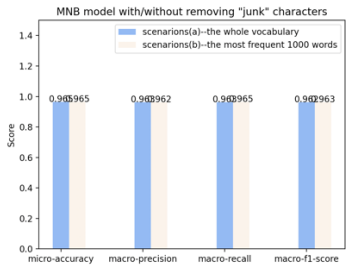


Micro	accuracy	precision	recall	F1
a	0.965	0.965	0.965	0.965
b	0.965	0.965	0.965	0.965

Macro:



Macro	accuracy	precision	recall	F1
a	0.97	0.905	0.855	0.867
b	0.91	0.913	0.899	0.904



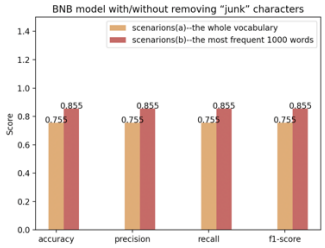
Micro	accuracy	precision	recall	F1
a	0.965	0.963	0.963	0.962
b	0.965	0.962	0.965	0.963

The BNB and MNB model by using the most 1000 words have higher figures. In addition, comparing with MNB model, BNB have lower figures.

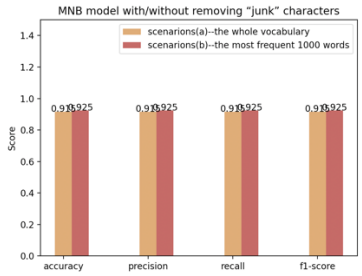
(ii)

The results of data from *summaries.tsv*

Micro:

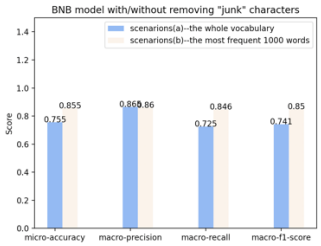


Micro	accuracy	precision	recall	F1
a	0.755	0.755	0.755	0.755
b	0.855	0.855	0.855	0.855

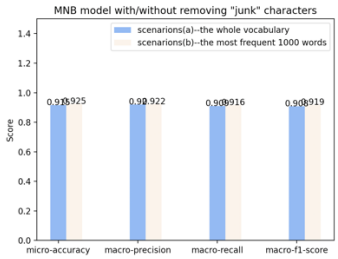


Micro	accuracy	precision	recall	F1
a	0.915	0.915	0.915	0.915
b	0.925	0.925	0.925	0.925

Macro:



Macro	accuracy	precision	recall	F1
a	0.755	0.865	0.725	0.741
b	0.855	0.86	0.846	0.85



Macro	accuracy	precision	recall	F1
a	0.915	0.92	0.909	0.908
b	0.925	0.922	0.916	0.919

The BNB and MNB model by using the most 1000 words have higher figures. In addition, comparing with MNB model, BNB have lower figures.

(iii)

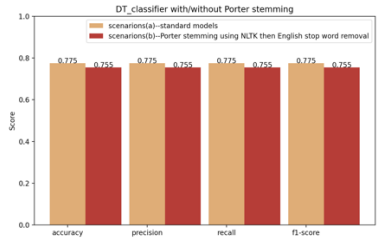
Comparing with (i), BNB and MNB model show a more unstable performance in sunmmarise.tsv. However, we can find that using the most frequent 1000 words is effective in improving all figures for BNB and MNB

3.

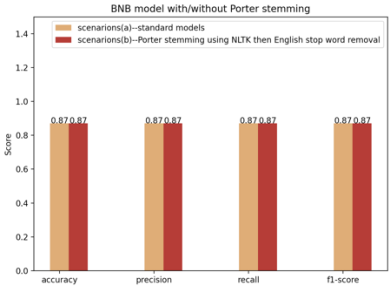
(i)

The results of data from *article.tsv*

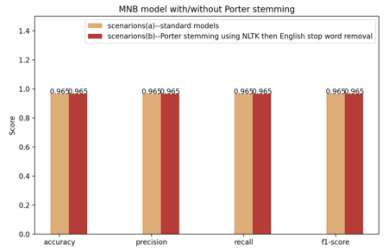
Micro:



Micro	accuracy	precision	recall	F1
a	0.775	0.775	0.775	0.775
b	0.755	0.755	0.755	0.755

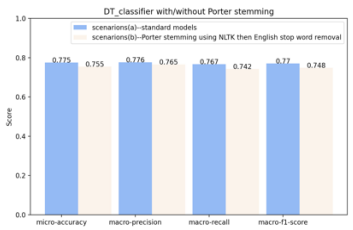


Micro	accuracy	precision	recall	F1
a	0.87	0.87	0.87	0.87
b	0.87	0.87	0.87	0.87

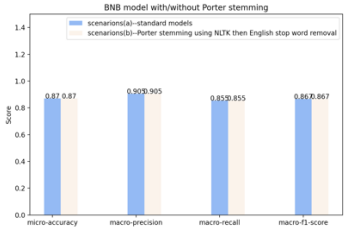


Micro	accuracy	precision	recall	F1
a	0.965	0.965	0.965	0.965
b	0.965	0.965	0.965	0.965

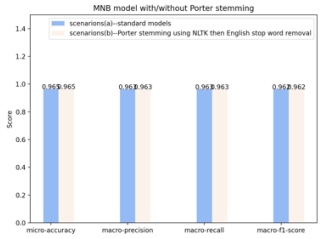
Macro:



Macro	accuracy	precision	recall	F1
a	0.775	0.776	0.767	0.77
b	0.755	0.765	0.742	0.748



Macro	accuracy	precision	recall	F1
a	0.87	0.905	0.855	0.867
b	0.87	0.905	0.855	0.867



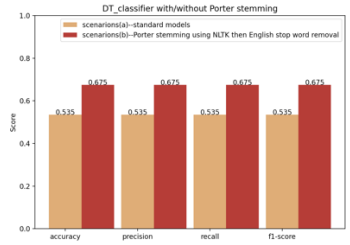
Macro	accuracy	precision	recall	F1
a	0.965	0.963	0.963	0.962
b	0.965	0.965	0.963	0.962

To compare the figures, we can find that Porter stemming seems not effective in improving the figures, and in DT model it even leads the figures lower.

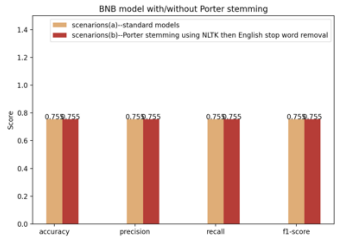
(ii)

The results of data from *summaries.tsv*

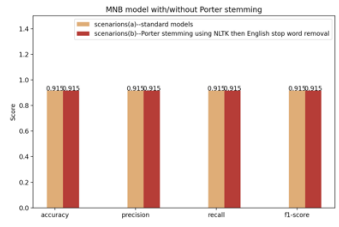
Micro:



Micro	accuracy	precision	recall	F1
a	0.535	0.535	0.535	0.535
b	0.675	0.675	0.675	0.675

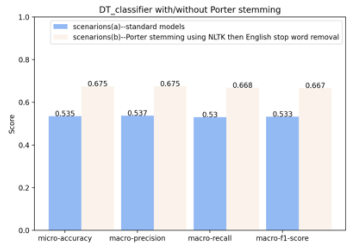


Micro	accuracy	precision	recall	F1
a	0.755	0.755	0.755	0.755
b	0.755	0.755	0.755	0.755

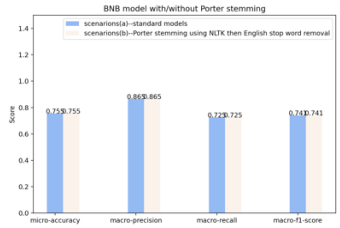


Micro	accuracy	precision	recall	F1
a	0.915	0.915	0.915	0.915
b	0.915	0.915	0.915	0.915

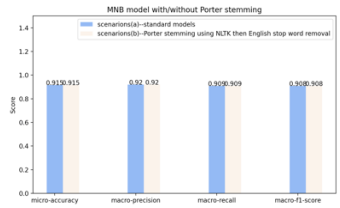
Macro:



Macro	accuracy	precision	recall	F1
a	0.535	0.537	0.53	0.533
b	0.675	0.675	0.668	0.667



Macro	accuracy	precision	recall	F1
a	0.755	0.865	0.725	0.741
b	0.755	0.865	0.725	0.741



Micro	accuracy	precision	recall	F1
a	0.915	0.92	0.909	0.908
b	0.915	0.92	0.909	0.908

In summarise.tsv, we can find that Porter stemming have no efficient to BNB model and MNB model. It only works in DT model and have significant improving.

(iii)

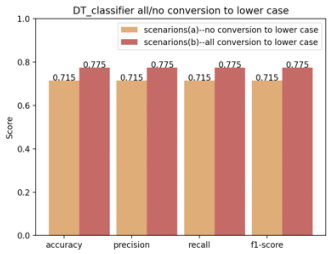
Comparing the results of (i) and (ii), we can find that Porter stemming have a little impact to the figure of BNB model and MNB model. It works in DT model, and with the difference of data, the result of DT is unpredictable.

4.

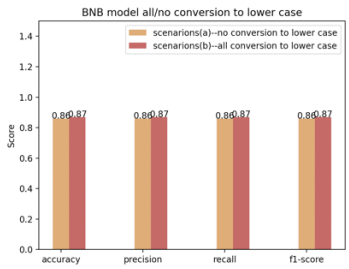
(i)

The results of data from *article.tsv*

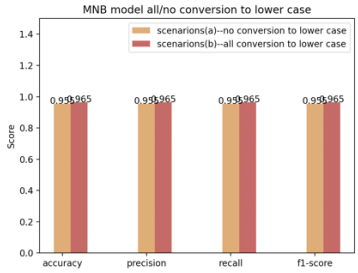
Micro:



Micro	accuracy	precision	recall	F1
a	0.715	0.715	0.715	0.715
b	0.775	0.775	0.775	0.775

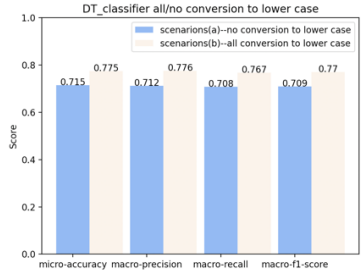


Micro	accuracy	precision	recall	F1
a	0.86	0.86	0.86	0.86
b	0.87	0.87	0.87	0.87

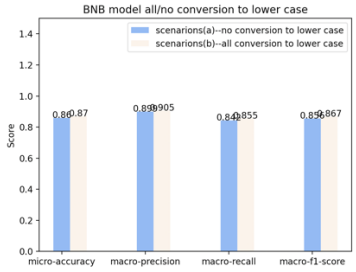


Micro	accuracy	precision	recall	F1
a	0.955	0.955	0.955	0.955
b	0.965	0.965	0.965	0.965

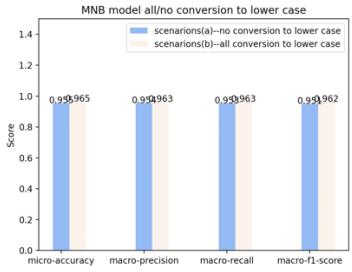
Macro:



Macro	accuracy	precision	recall	F1
a	0.715	0.712	0.708	0.709
b	0.775	0.776	0.767	0.77



Macro	accuracy	precision	recall	F1
a	0.86	0.899	0.842	0.856
b	0.87	0.905	0.855	0.867



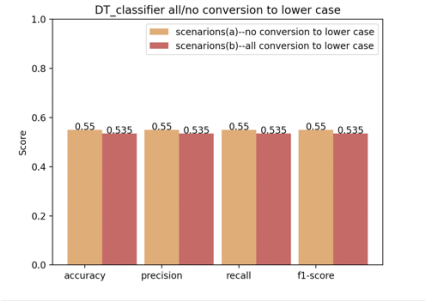
Macro	accuracy	precision	recall	F1
a	0.955	0.954	0.953	0.951
b	0.965	0.963	0.963	0.962

Comparing the figure above, we can find that lower case can improve all the figures in each model.

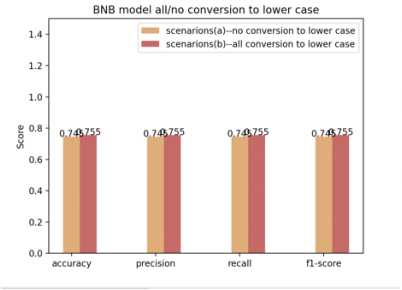
(ii)

The results of data from *summaries.tsv*

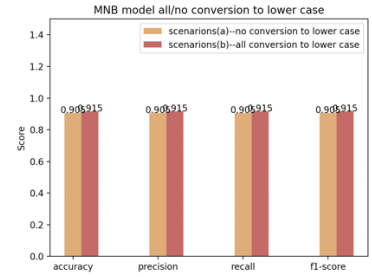
Micro:



Micro	accuracy	precision	recall	F1
a	0.55	0.55	0.55	0.55
b	0.535	0.535	0.535	0.535

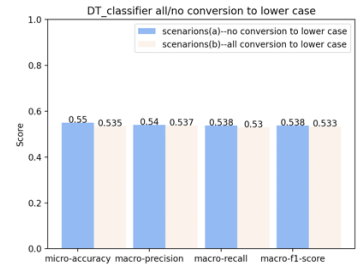


Micro	accuracy	precision	recall	F1
a	0.915	0.92	0.909	0.908
b	0.925	0.922	0.916	0.919

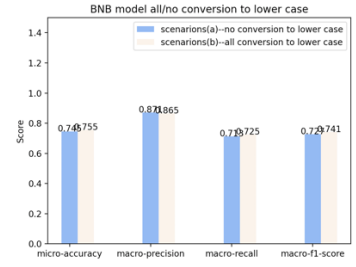


Micro	accuracy	precision	recall	F1
a	0.905	0.905	0.905	0.905
b	0.915	0.915	0.915	0.915

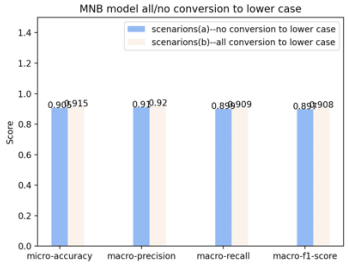
Macro:



Micro	accuracy	precision	recall	F1
a	0.55	0.54	0.538	0.538
b	0.535	0.537	0.53	0.533



Micro	accuracy	precision	recall	F1
a	0.745	0.871	0.713	0.727
b	0.755	0.865	0.725	0.741



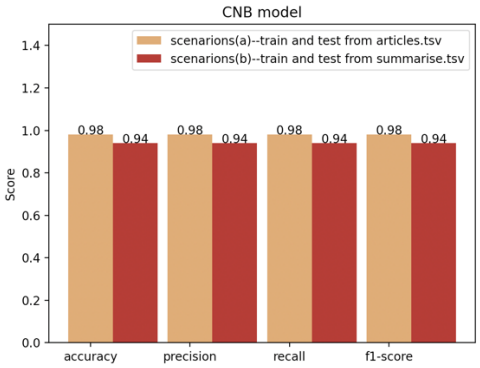
Micro	accuracy	precision	recall	F1
a	0.905	0.91	0.899	0.897
b	0.915	0.92	0.909	0.908

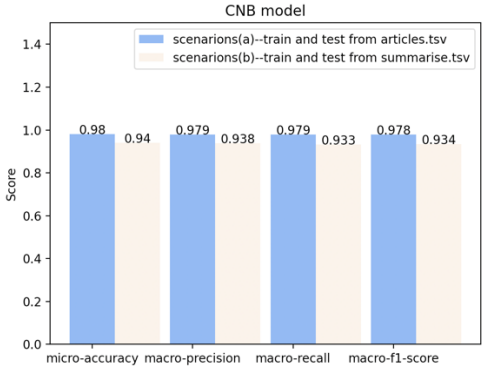
We can find that lower case can improve most of figures except in DT, and DT model also shows the lowest figures – nearly half of the accuracy, precision, recall and F1.

(iii)

To compare with (i) and (ii), DT is the worst and the most unstable model in these models. Opposite, MNB can always shows higher figures in both of the data from articles.tsv and summarises.tsv

5.





In CNB model, I try to use the most frequent 1000 words and use the same preprocessing with other models. With training and testing data by CNB, I got higher figures than by DT, BNB and MNB.