BOĞAZİÇİ UNIVERSITY

CMPE 493

ASSIGNMENT 2

Spring 2018

Text Classification using Naive Bayes

Mustafa Enes ÇAKIR

April 3, 2018

1 Document Counts

1.1 Train

- *earn:* 2848
- *acq:* 1617
- *money-fx:* 536
- *grain:* 429
- *crude:* 361
- *Total:* 5791

1.2 Test

- *earn:* 1084
- *acq:* 710
- *money-fx:* 178
- *grain:* 148
- *crude:* 180
- *Total:* 2300

2 Selected Features

2.1 earn

vs, cts, shr, net, said, qtr, to, revs, lt, note, loss, s, 4th, profit, has, u, at, div, dividend, 31, record, 1, avg, shrs, qtly, prior, year, pct, not, would, mths, market, agreement, offer, had, buy, oper, agreed, exchange, 2, acquire, today, about, 1st, were, bank, more, pay, jan, payout

2.2 acq

cts, said, shr, net, qtr, shares, acquisition, acquire, to, stake, note, year, company, merger, offer, loss, has, buy, record, 1, terms, common, unit, transaction, acquired, group, corp, profit, sell, purchase, inc, outstanding, undisclosed, bid, completed, agreed, agreement, 31, stock, shrs, takeover, pct, lt, dividend, commission, completes, subsidiary, disclosed, approval, investor

2.3 money-fx

lt, bank, dollar, money, market, currency, central, rates, treasury, dealers, rate, yen, cts, england, u, currencies, japan, inc, monetary, around, paris, bills, shortage, nations, intervention, today, net, company, fed, at, exchange, assistance, corp, foreign, 000, deficit, system, k, economic, liquidity, dollars, reserve, banks, germany, stg, trade, further, share, accord, against

2.4 grain

wheat, tonnes, lt, agriculture, grain, corn, usd, export, crop, u, department, farmers, cts, vs, s, soviet, inc, crops, tonne, farm, maize, barley, net, program, agricultural, commodity, grains, enhancement, rice, company, ussr, ec, feed, soybeans, winter, soybean, corp, season, union, trade, 87, growers, note, sorghum, to, shipment, commodities, cereals, bushel, hectares

2.5 crude

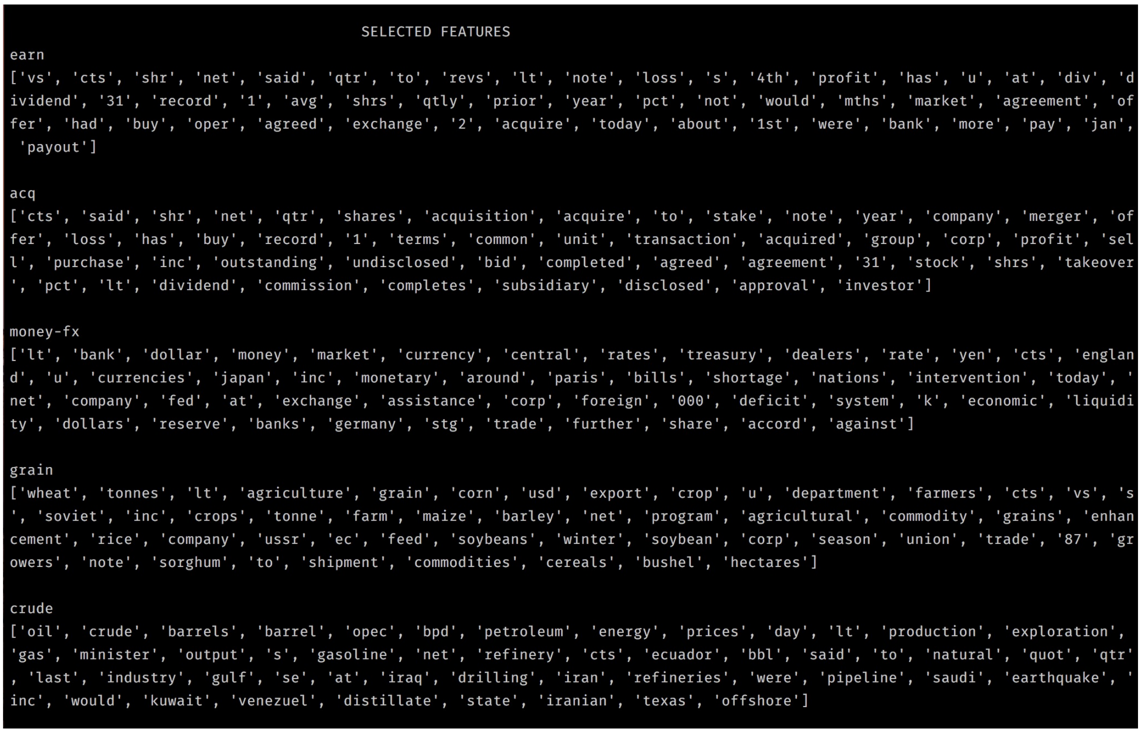
oil, crude, barrels, barrel, opec, bpd, petroleum, energy, prices, day, lt, production, exploration, gas, minister, output, s, gasoline, net, refinery, cts, ecuador, bbl, said, to, natural, quot, qtr, last, industry, gulf, se, at, iraq, drilling, iran, refineries, were, pipeline, saudi, earthquake, inc, would, kuwait, venezuel, distillate, state, iranian, texas, offshore

3 Performance Values

Table 1: Performance Values

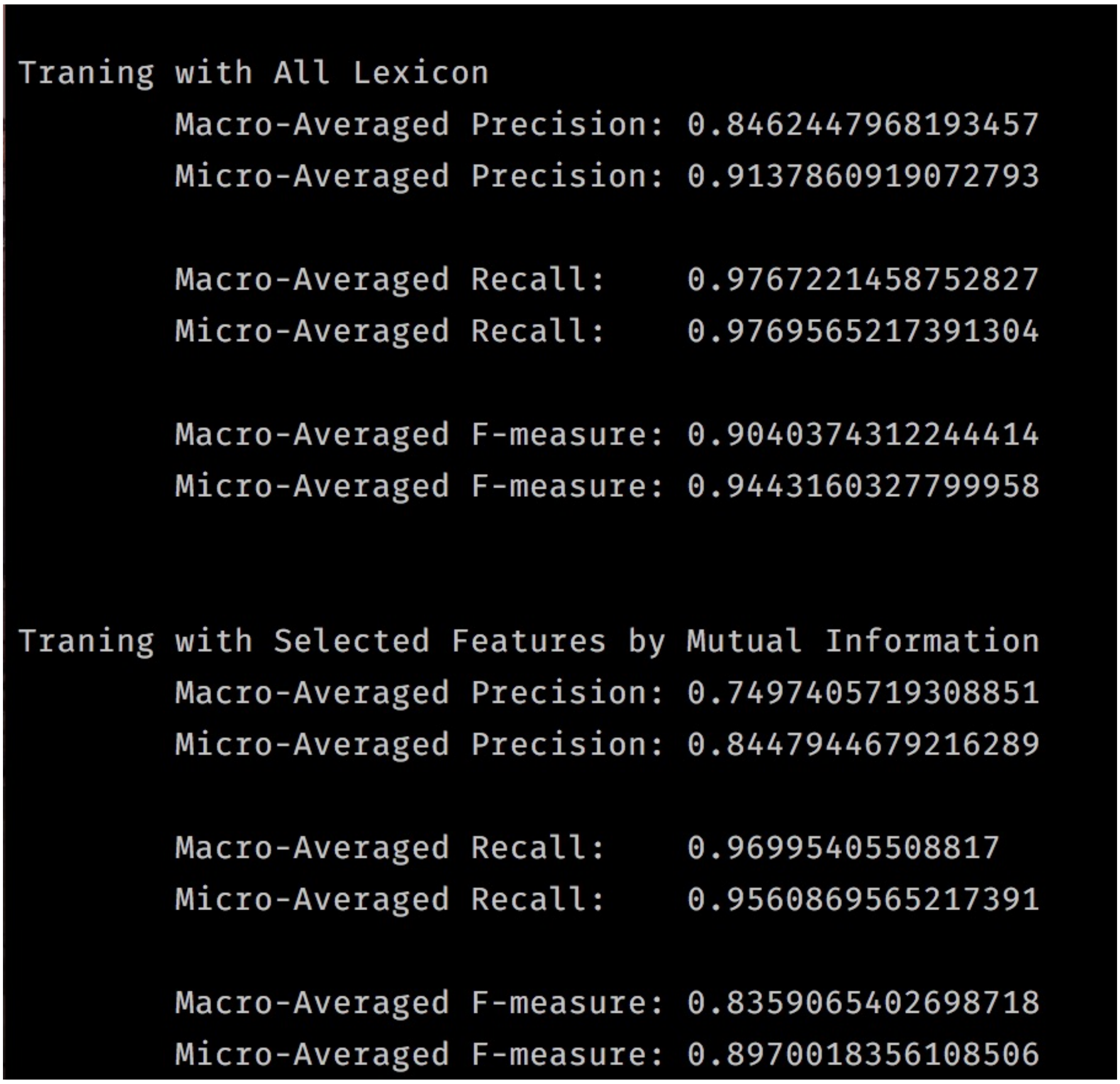
	All Lexicon			Mutual Information		
Class	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
earn	0.978	0.973	0.975	0.974	0.932	0.952
acq	0.945	0.983	0.963	0.892	0.976	932
money-fx	0.773	0.994	0.870	0.639	0.994	0.778
grain	0.749	0.966	0.844	0.606	0.986	0.743
crude	0.787	0.967	0.868	0.648	0.961	0.774
Macro Ave.	0.846	0.977	0.904	0.751	0.970	0.836
Micro Ave.	0.913	0.977	0.944	0.845	0.956	0.897

4 Screenshots



```
                              SELECTED FEATURES
earn
['vs', 'cts', 'shr', 'net', 'said', 'qtr', 'to', 'revs', 'lt', 'note', 'loss', 's', '4th', 'profit', 'has', 'u', 'at', 'div', 'd
ividend', '31', 'record', '1', 'avg', 'shrs', 'qtly', 'prior', 'year', 'pct', 'not', 'would', 'mths', 'market', 'agreement', 'of
fer', 'had', 'buy', 'oper', 'agreed', 'exchange', '2', 'acquire', 'today', 'about', '1st', 'were', 'bank', 'more', 'pay', 'jan',
 'payout']

acq
['cts', 'said', 'shr', 'net', 'qtr', 'shares', 'acquisition', 'acquire', 'to', 'stake', 'note', 'year', 'company', 'merger', 'of
fer', 'loss', 'has', 'buy', 'record', '1', 'terms', 'common', 'unit', 'transaction', 'acquired', 'group', 'corp', 'profit', 'sel
l', 'purchase', 'inc', 'outstanding', 'undisclosed', 'bid', 'completed', 'agreed', 'agreement', '31', 'stock', 'shrs', 'takeover
', 'pct', 'lt', 'dividend', 'commission', 'completes', 'subsidiary', 'disclosed', 'approval', 'investor']

money-fx
['lt', 'bank', 'dollar', 'money', 'market', 'currency', 'central', 'rates', 'treasury', 'dealers', 'rate', 'yen', 'cts', 'englan
d', 'u', 'currencies', 'japan', 'inc', 'monetary', 'around', 'paris', 'bills', 'shortage', 'nations', 'intervention', 'today', '
net', 'company', 'fed', 'at', 'exchange', 'assistance', 'corp', 'foreign', '000', 'deficit', 'system', 'k', 'economic', 'liquidi
ty', 'dollars', 'reserve', 'banks', 'germany', 'stg', 'trade', 'further', 'share', 'accord', 'against']

grain
['wheat', 'tonnes', 'lt', 'agriculture', 'grain', 'corn', 'usd', 'export', 'crop', 'u', 'department', 'farmers', 'cts', 'vs', 's
', 'soviet', 'inc', 'crops', 'tonne', 'farm', 'maize', 'barley', 'net', 'program', 'agricultural', 'commodity', 'grains', 'enhan
cement', 'rice', 'company', 'ussr', 'ec', 'feed', 'soybeans', 'winter', 'soybean', 'corp', 'season', 'union', 'trade', '87', 'gr
owers', 'note', 'sorghum', 'to', 'shipment', 'commodities', 'cereals', 'bushel', 'hectares']

crude
['oil', 'crude', 'barrels', 'barrel', 'opec', 'bpd', 'petroleum', 'energy', 'prices', 'day', 'lt', 'production', 'exploration',
'gas', 'minister', 'output', 's', 'gasoline', 'net', 'refinery', 'cts', 'ecuador', 'bbl', 'said', 'to', 'natural', 'quot', 'qtr'
, 'last', 'industry', 'gulf', 'se', 'at', 'iraq', 'drilling', 'iran', 'refineries', 'were', 'pipeline', 'saudi', 'earthquake', '
inc', 'would', 'kuwait', 'venezuel', 'distillate', 'state', 'iranian', 'texas', 'offshore']
```

Figure 1: Selected Features



```
Traning with All Lexicon
        Macro-Averaged Precision: 0.8462447968193457
        Micro-Averaged Precision: 0.9137860919072793

        Macro-Averaged Recall:    0.9767221458752827
        Micro-Averaged Recall:    0.9769565217391304

        Macro-Averaged F-measure: 0.9040374312244414
        Micro-Averaged F-measure: 0.9443160327799958


Traning with Selected Features by Mutual Information
        Macro-Averaged Precision: 0.7497405719308851
        Micro-Averaged Precision: 0.8447944679216289

        Macro-Averaged Recall:    0.96995405508817
        Micro-Averaged Recall:    0.9560869565217391

        Macro-Averaged F-measure: 0.8359065402698718
        Micro-Averaged F-measure: 0.8970018356108506
```

Figure 2: Results