



Novel Balanced Feature Representation for Wikipedia Vandalism Detection Task

István Hegedűs, Róbert Ormándi,
Richárd Farkas, and Márk Jelasity

University of Szeged

Hungary

ihegedus@inf.u-szeged.hu



Our approach

- Supervised learning
- Rich feature set
- Meta-learning scheme

Vector space model (VSM)

- unigrams
- values:
 - N if does not occur in the edit
 - A if in added sequence
 - D if in removed sequence
 - C if in changed sequence
- #features = 47 324
- best 100 by InfoGain



Balanced VSM

- sample is unbalanced
 - 93.9% regular

- BVSM:

```
for i in 1 to N do
```

```
    D = vandalism AND random_regular
```

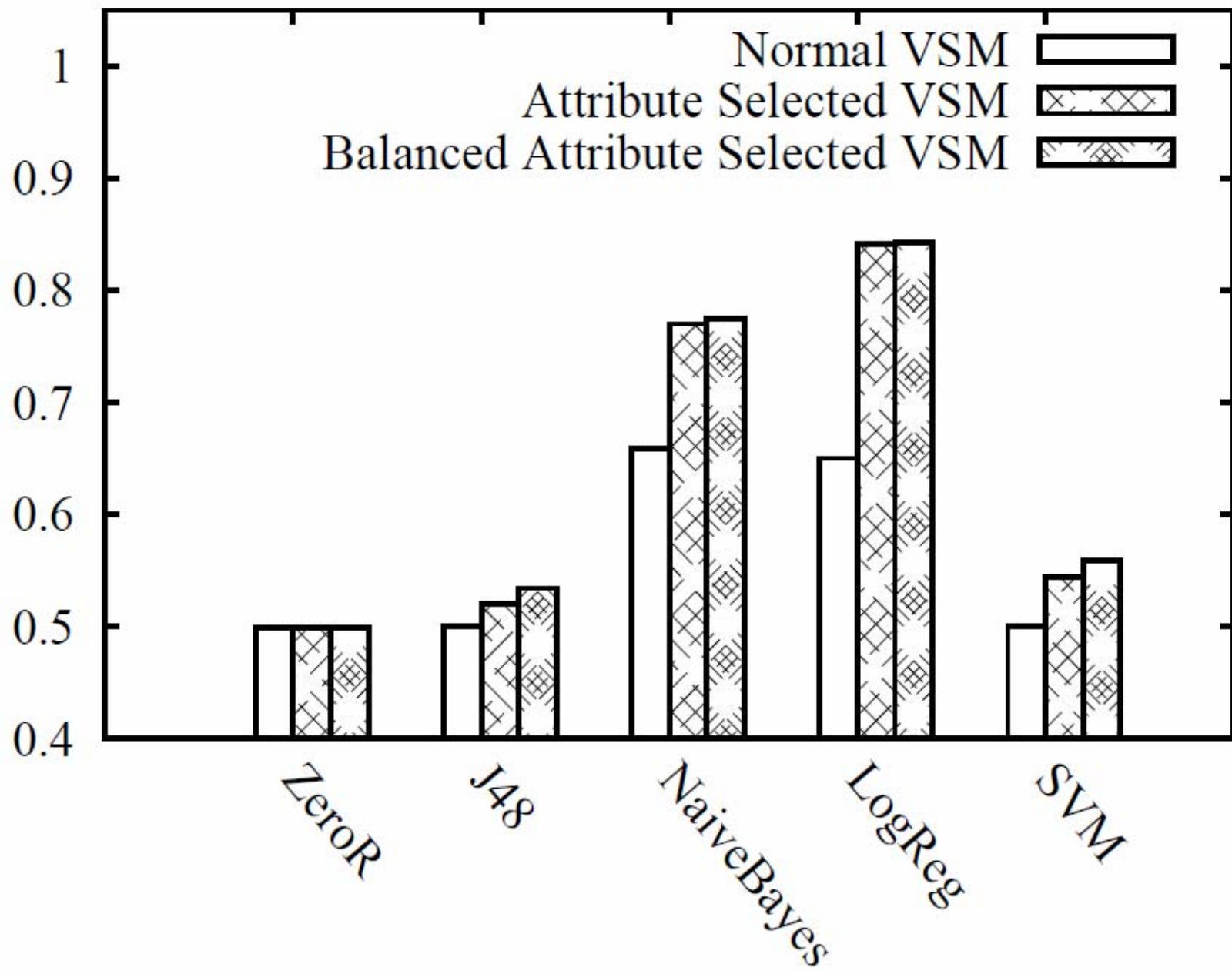
```
    IG += InfoGainScore(D)
```

```
done
```

```
VSM = best(IG, 100)
```



10-fold AUC(Vandalism)





Other features

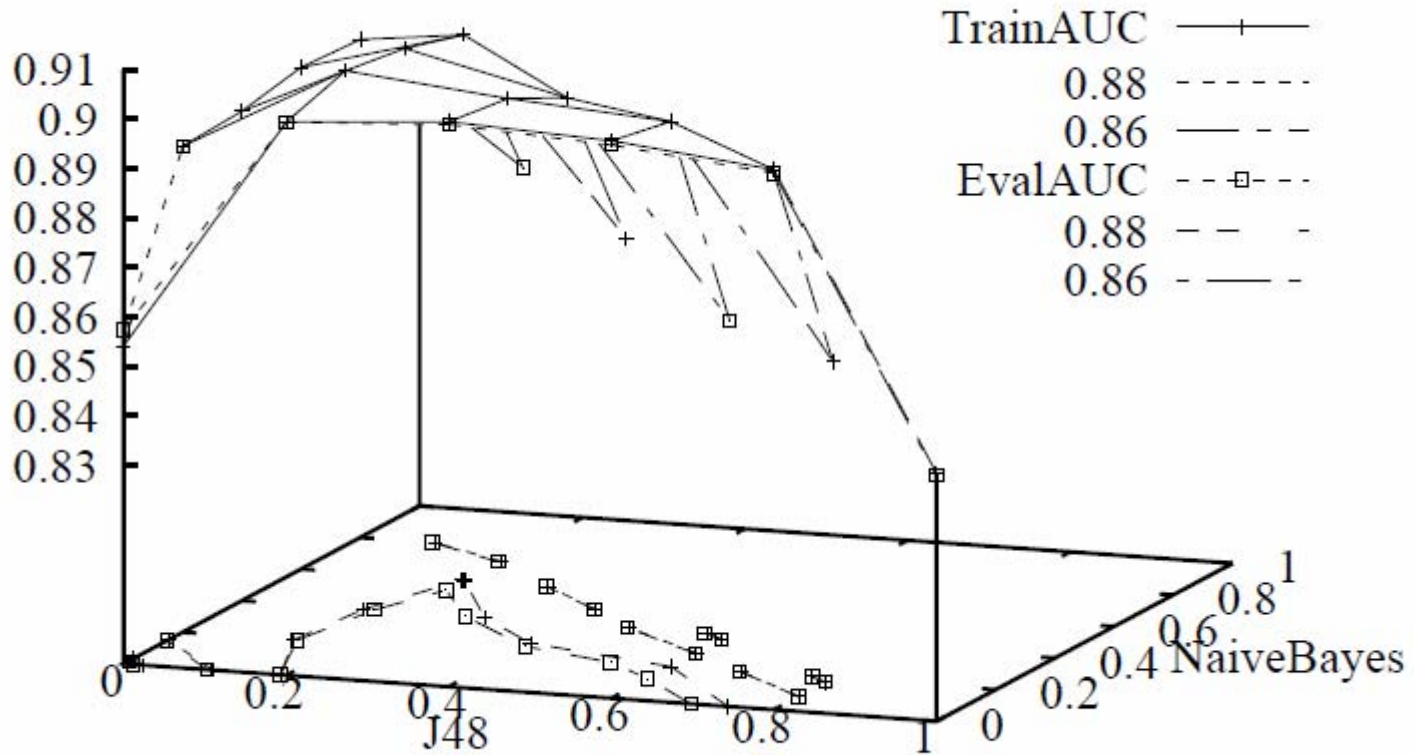
- CharacterStatistic
 - uppercase and lowercase ratio
- RepeatedCharSequences
 - asdasdasdasdasd
- ValidWordRatio
 - English/pejorative words
- CommentStatistic
- UserNameOrIP
 - nickname or country from IP

10-fold-cross-validation

	AUC (10-fold)
Balanced VSM	0.813
Balanced VSM + stopword	0.843
Other features	0.883
Other + unbalanced VSM	0.884
Other + balanced VSM	0.887



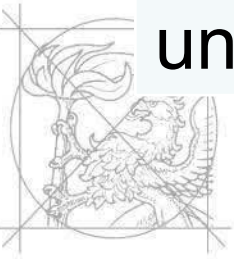
Meta learning



J48=0.3; NaiveBayes=0.09; Logistic=0.61

Results (eval)

	AUC (LogReg)	AUC (Voting)
Balanced VSM	0.744	0.761
Other features	0.865	0.876
Other + balanced	0.854	0.877
Other + unbalanced	0.864	0.880





Summary

- VSM has no significant added value
- meta-learning (+2%)

