Cross Language Prediction of Vandalism on Wikipedia using Article Views and Revisions

Khoi-Nguyen Tran and Peter Christen

Research School of Computer Science
The Australian National University
{khoi-nguyen.tran, peter.christen}@anu.edu.au

PAKDD 2013 – Gold Coast, Australia

Introduction

- Vandalism is a major issue on Wikipedia
  – 2% of revisions (identified by contributors)
- Traces of malicious behaviour in
  – Edit logs (revisions data set)
  – Access logs (pagecounts data set)
- Why cross language?
  – Relatively fewer contributors for non-English
  – Variety of vandalism patterns

Wikipedia Data Sets

- Languages: English, German
- Revision Data Set
  - Captures all edits made
  - Commonly used for research
- Article Views (per hour)
  - Count of requests per article for each hour
  - Available since Dec 2007
  - Few research papers use this data

Vandalised Revisions

English

- Articles: ~4 million
- Revisions: ~305 million
- All users: ~4 million
- All IPs: ~25 million

German

- Articles: ~1.4 million
- Revisions: ~65 million
- All users: ~0.4 million
- All IPs: ~5 million

- First data dump of June 2012
- Vandalism identified in revision comment
  - E.g. "… **rev(ert)** … **vandalism** …" , "**rvv**", etc.
- Choose language independent features

Article Views

English

- Articles: ~2.2 million
- Total: ~4,500 million

German

- Articles: ~0.8 million
- Total: ~1,500 million

- From January 2012 to May 2012
- View counts aggregated for each hour
- Matched with revisions to label views of vandalism
  - This shows “exposure” of vandalised revisions
- Used all available features

Combined Data Set

View Features

- Project name
- Article Title ⊗
- Hour timestamp ⊗

O Number of requests ⊗

O Bytes transferred ⊗

Revision Features

- Article title
- Hour timestamp

X Anonymous edit ⊗

X Minor revision ⊗

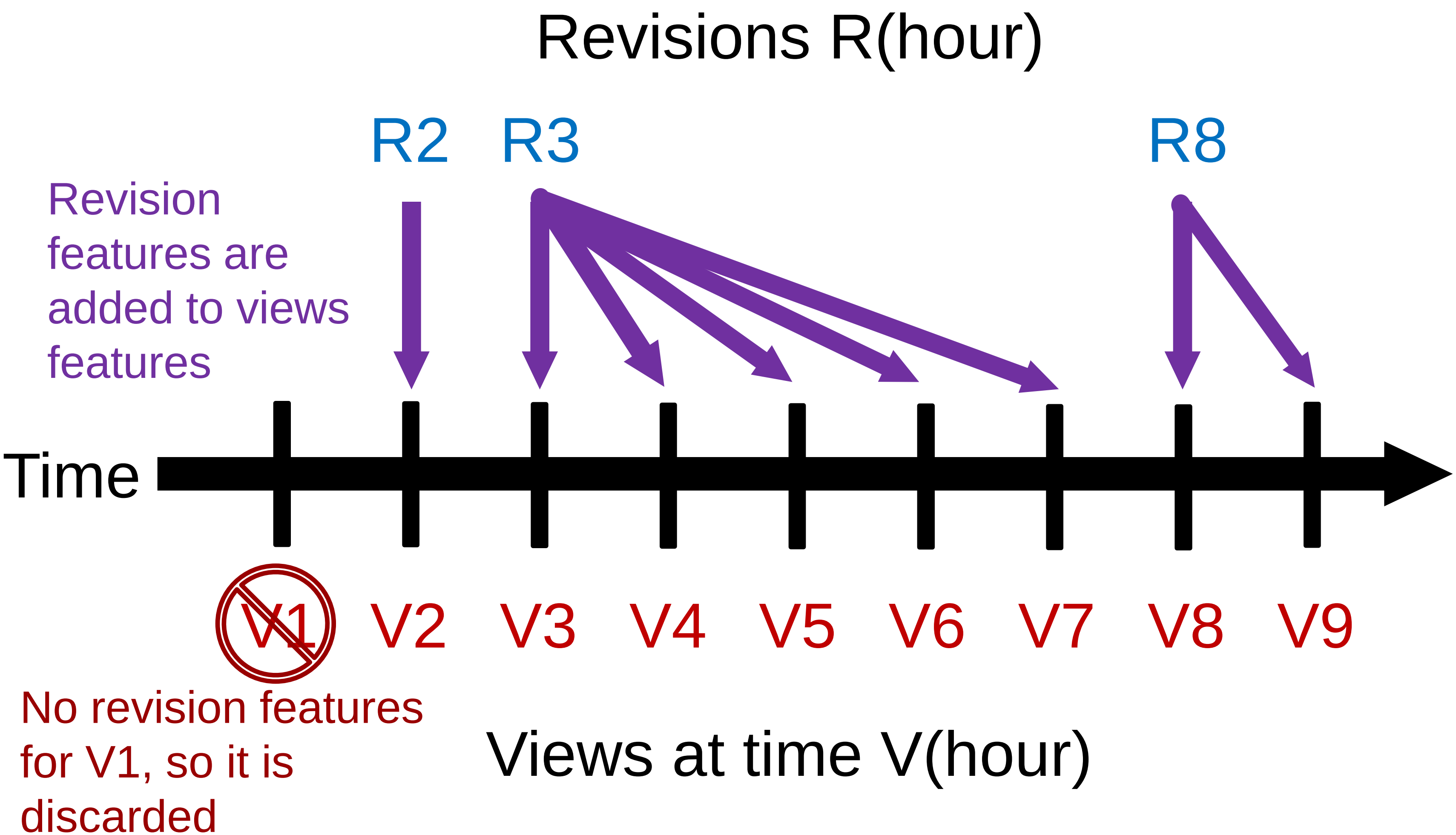
X Size of comment ⊗

X Size of article text ⊗

X Vandalism (class label) ⊗

- Combined Features ⊗ - Exposure to vandalism

Combined Data Set - Visualisation



Combined Data Set

- Timespan of data sets
  - Training set: January to April 2012
  - Testing set: May 2012

English

- Train (Combined)
  - Views: ~270 million
  - Vandalised: ~6 million
- Test (Combined)
  - Views: ~100 million
  - Vandalised: ~2 million

German

- Train (Combined)
  - Views: ~140 million
  - Vandalised: ~85,000
- Test (Combined)
  - Views: ~55 million
  - Vandalised: ~40,000

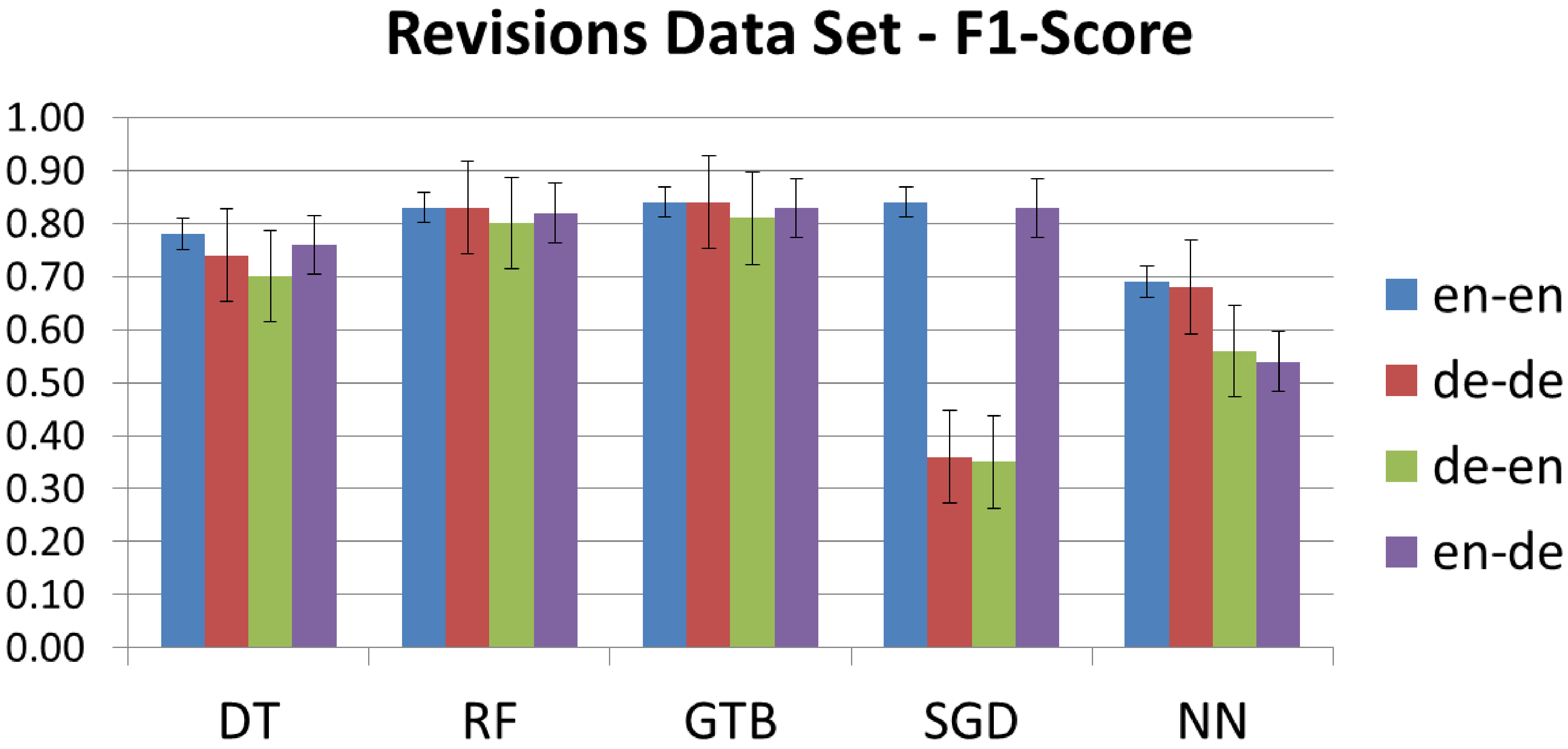
Experimental Setup

- Balanced data sets with under sampling
  - English
    - Training set: ~6 million samples of each class
    - Testing set: ~2 million samples of each class
  - German
    - Training set: ~85,000 samples of each class
    - Testing set: ~40,000 samples of each class
- Selected features of both data sets (revisions, views); all features (combined)

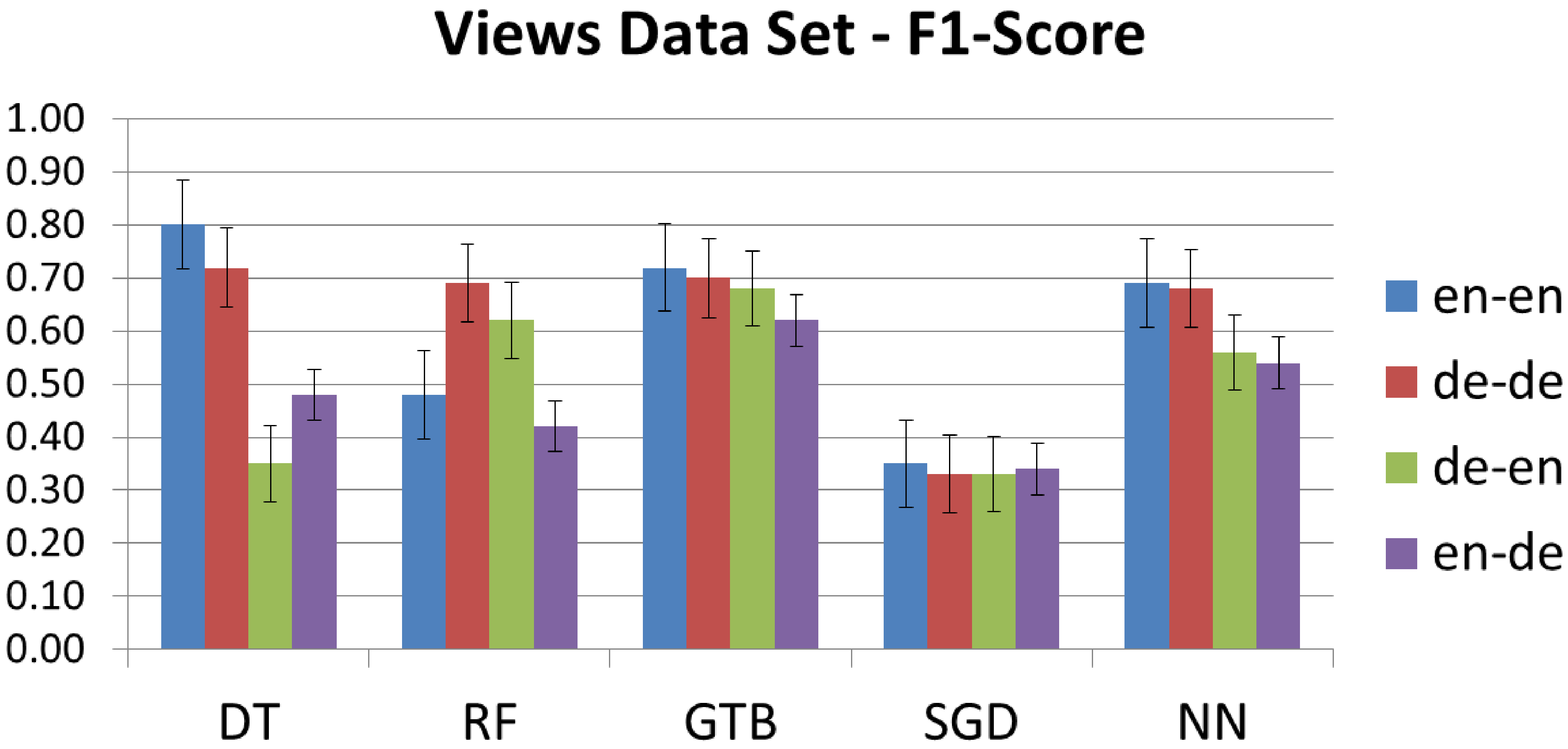
Experimental Setup

- Train models in one language, then classify on the other language (e.g. en-de)
- Scikit-learn toolkit
  - Decision Trees (DT)
  - Random Forest (RF)
  - Gradient Tree Boosting (GTB)
  - Stochastic Gradient Descent (SGD)
  - Nearest Neighbour (NN)

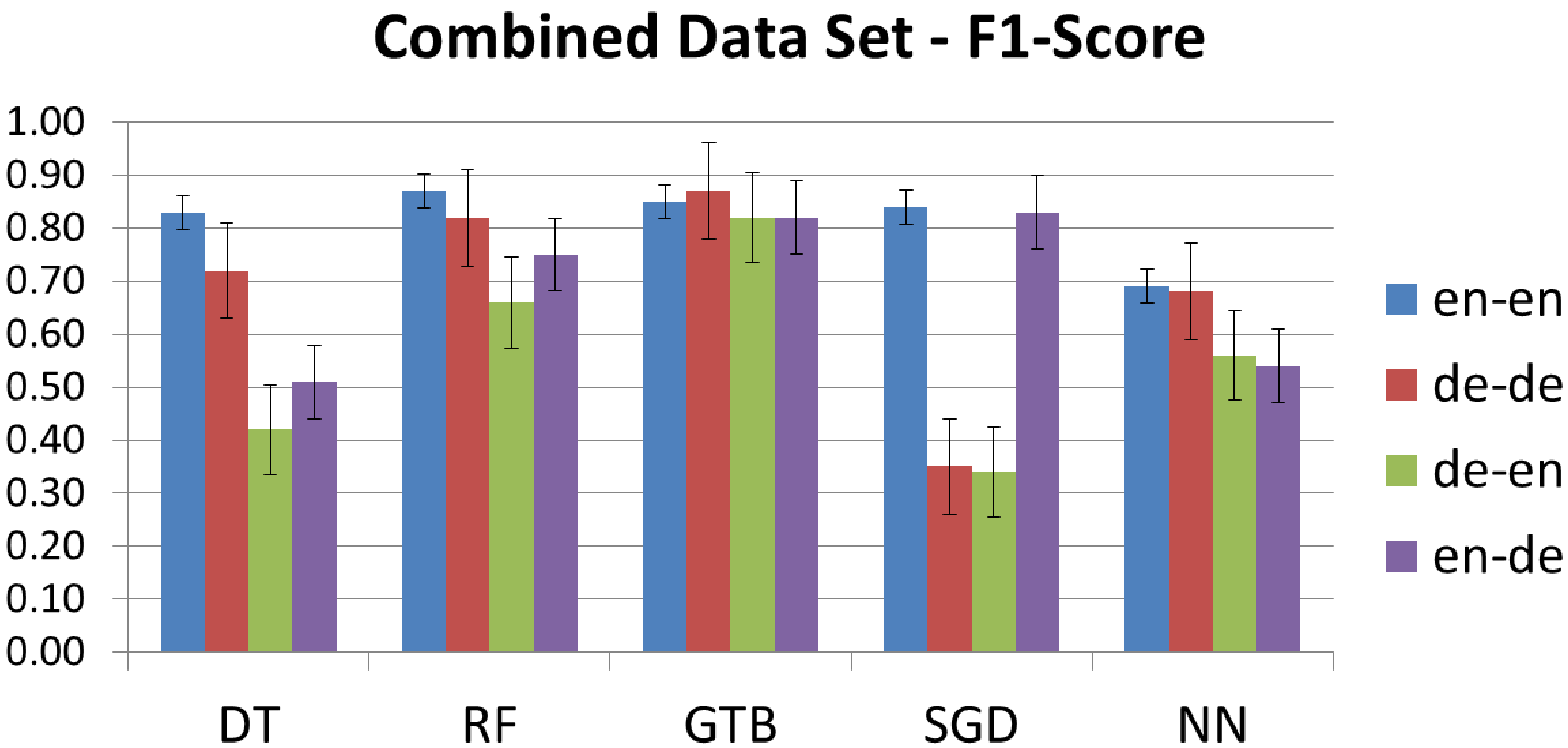
Experimental Results



Experimental Results



Experimental Results



Discussion

- Vandalism may have low invariants in behaviour across languages
- Cross Language Advantages
  - No significant loss in prediction quality when applied to another language
  - Useful for languages with few contributors and identified cases of vandalism
  - Article views data set is relatively simpler and provides reasonable prediction quality

Discussion

- Limitations
  - Few features considered
  - No analysis of revision content
  - Few types of classifier
  - Size of combined data set may be much larger than necessary
- Article views data sets may offer new vandalism patterns

Conclusion

- Demonstrated cross language prediction of vandalism
- Language independent features
- Developed 3 data sets
  - Article revisions
  - Article views
  - Combined

Conclusion

- Within the same language
  - High results: 87% F1-score
- Across languages
  - High results: 83% F1-score
- No significant loss of prediction quality across languages
- Gradient Tree Boosting showed generally best performance, but time consuming

Future Work

- Expand timespan of data set
- Apply to more languages
- Use more features from selection and generation
- Use other data balancing techniques
- Use this technique to feed results into more complex text based detectors

Thank You!

Conclusions

- Models trained in one language applied to another language
- High prediction results, but small loss in prediction quality

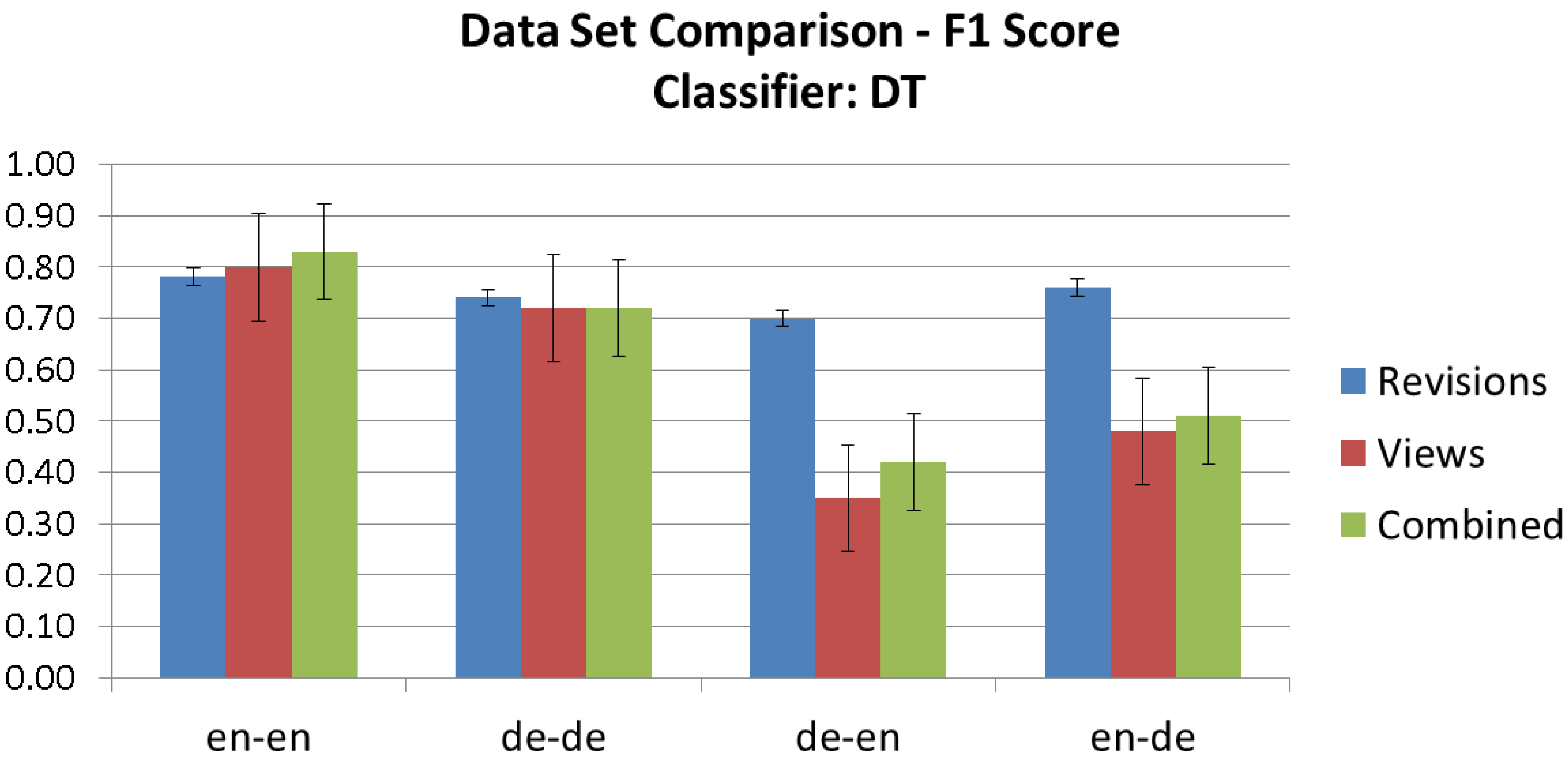
Future Work

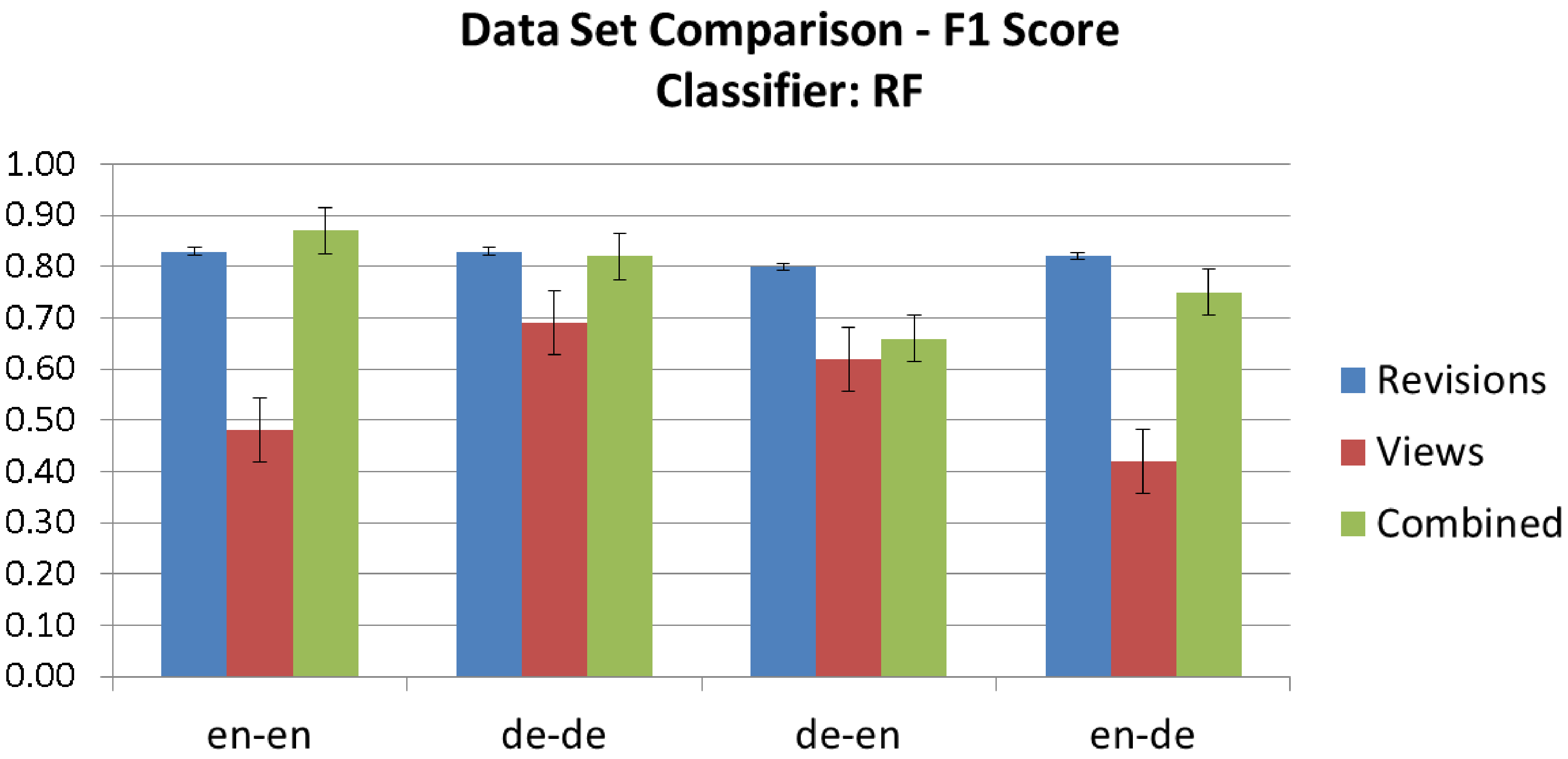
- More timespans, languages, features
- Other data balancing techniques
- Combine this technique with text based detectors

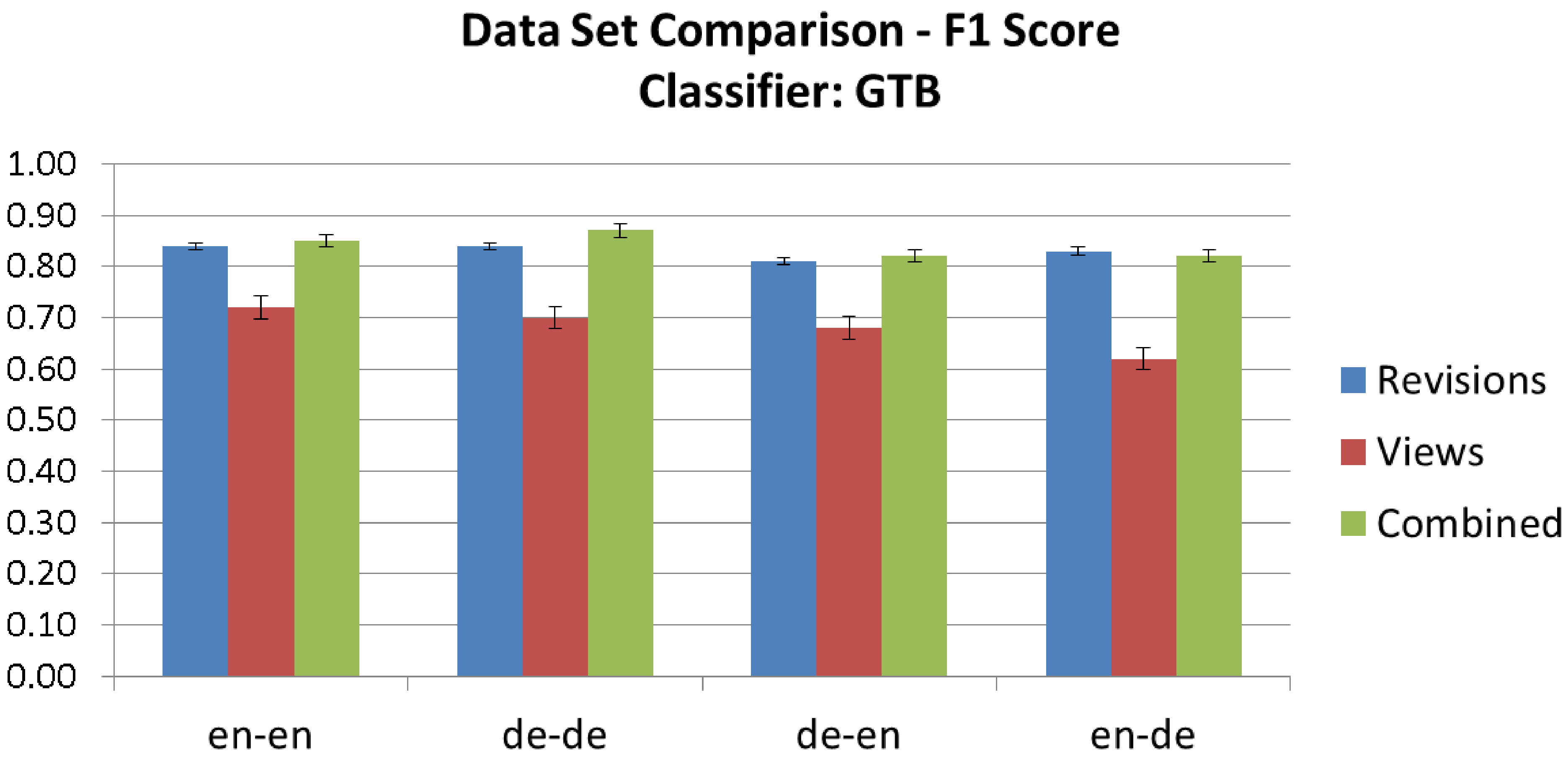
Contact: kndtran@cs.anu.edu.au

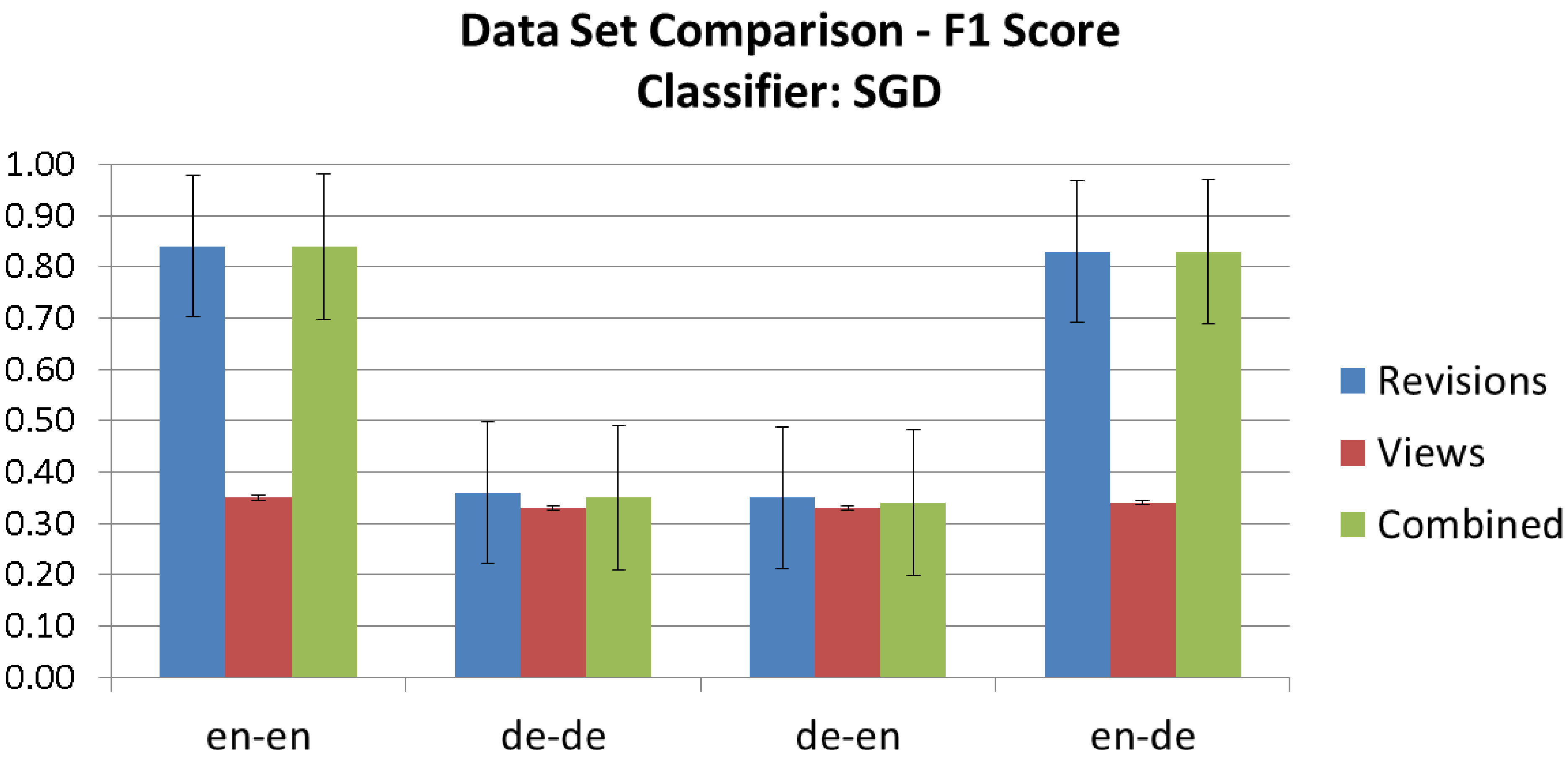
More info: kndtran.com

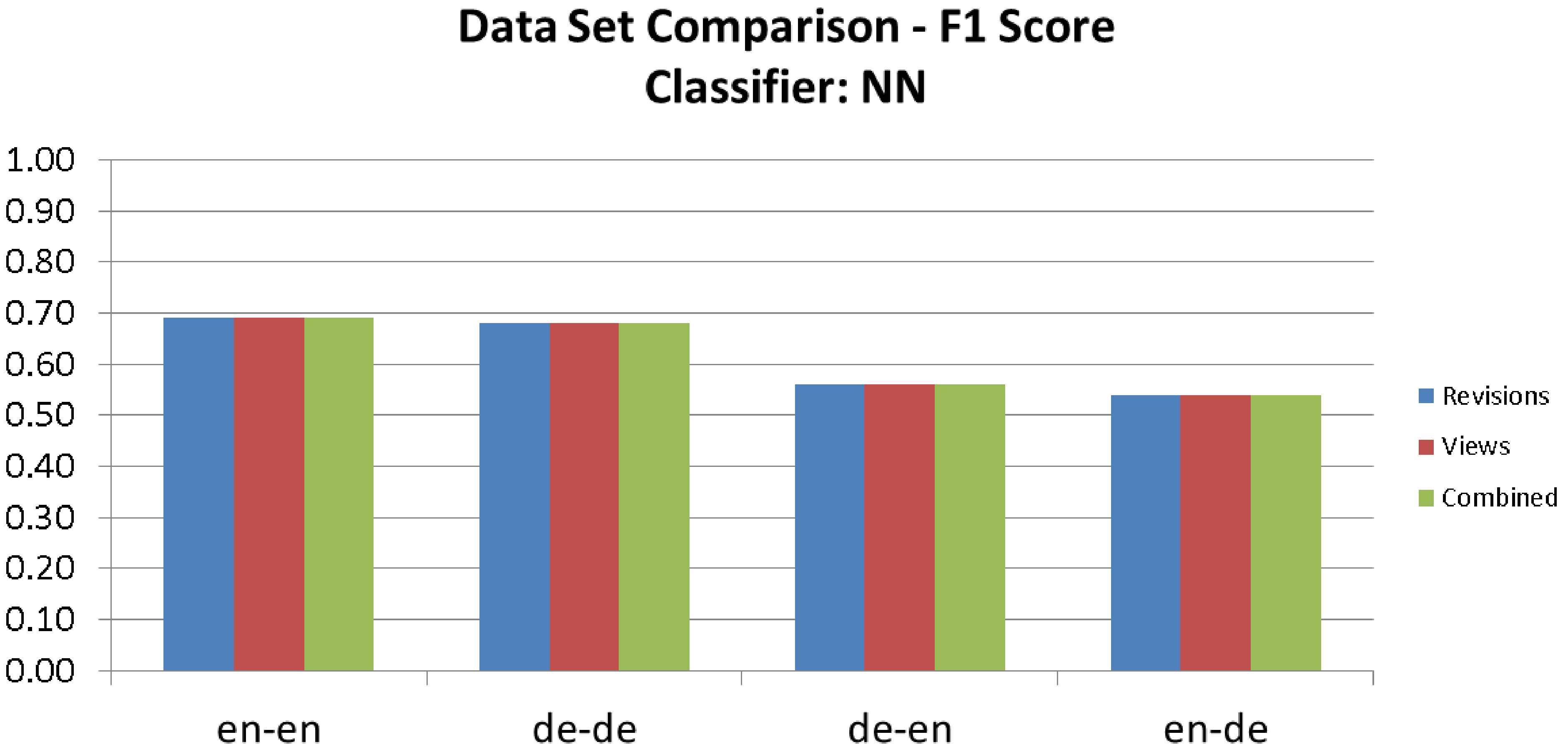
EXTRA SLIDES


Data Set Comparison - F1 Score
Classifier: DT
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
en-en
de-de
de-en
en-de
Revisions
Views
Combined


Data Set Comparison - F1 Score
Classifier: RF
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
en-en
de-de
de-en
en-de
Revisions
Views
Combined


Data Set Comparison - F1 Score
Classifier: GTB
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
en-en
de-de
de-en
en-de
Revisions
Views
Combined


Data Set Comparison - F1 Score
Classifier: SGD
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
en-en
de-de
de-en
en-de
Revisions
Views
Combined


Data Set Comparison - F1 Score
Classifier: NN
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
en-en
de-de
de-en
en-de
Revisions
Views
Combined

Combined Data Set - Visualisation

— T=4 O4 O4
— —5 O5 O5
— —6 O6 O6
— —7 O7 O7
— —8 O8 O8
— —9 O9 O9

— —2 X2 X2 X2 X2
— —5 X5 X5 X5 X5
— —8 X9 X9 X9 X9

Thus, we have the combined features:

— —4 O4 O4 X2 X2 X2 X2
— —5 O5 O5 X5 X5 X5 X5
— —6 O6 O6 X5 X5 X5 X5
— —7 O7 O7 X5 X5 X5 X5
— —8 O8 O8 X9 X9 X9 X9
— —9 O9 O9 X9 X9 X9 X9