4N – NEGATIVE NEWS NEURAL NETS

Dan Pavlovič

Darya Pisetskaya

Project Owner

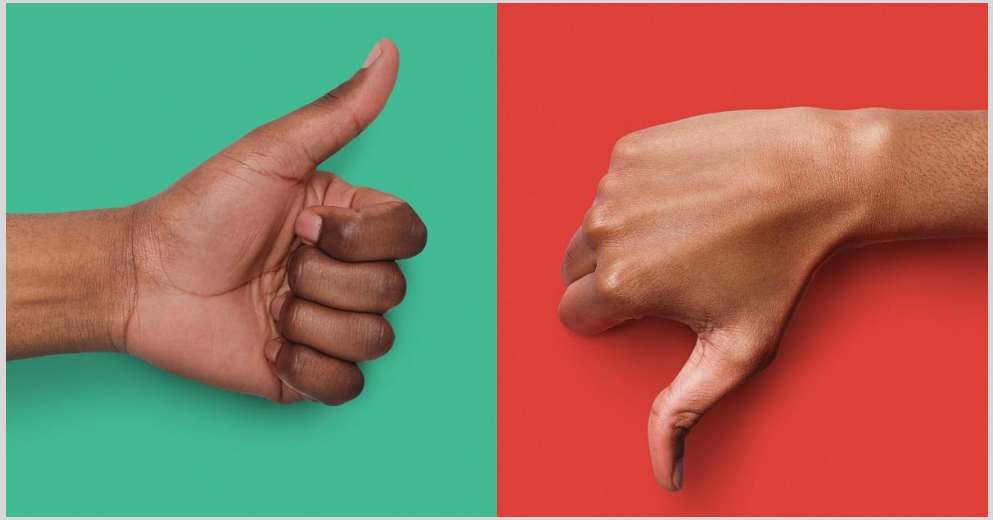
- Kristjan Roosild (Transfer Wise)

Project Description - Problem

- Financial organizations do compliance investigation on customers
- They look for adverse media
- Outsource or in-house
- **Manual check takes a long time**

Project Description - Solution

- Detect if article is adverse media or not
  - Adverse media is crime or crime suspicion news/article about a person or company (corruption, tax evasion, bribery, connection to terrorism, …)
- **Binary classification**



Data

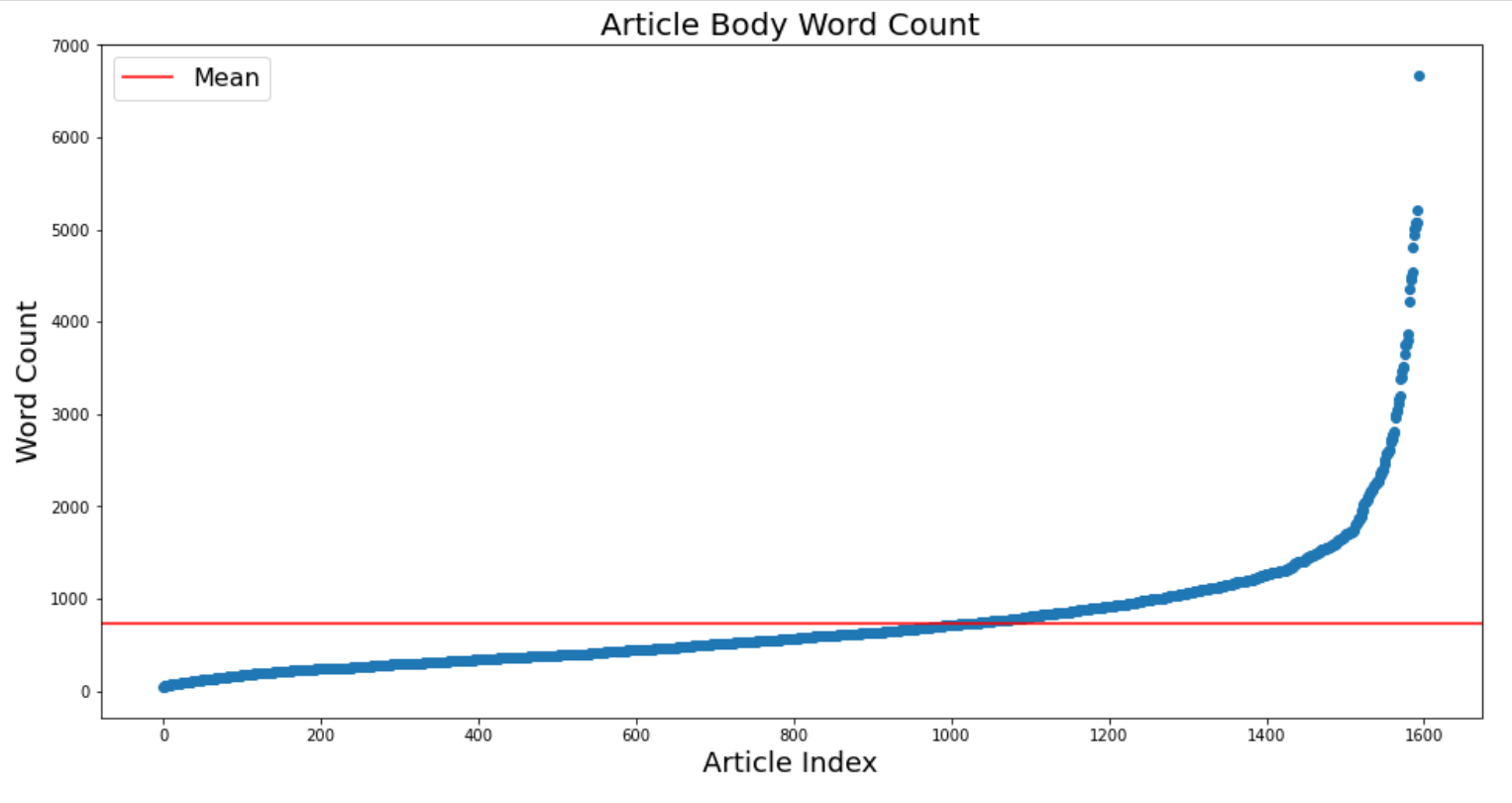
- For training:
  - 801 adverse media (AM) – label 1
  - 397 non-adverse media (NAM) – label 0
  - 396 random – label 0

title	article	label
Fridley Man Scams Thousands In Fraudulent Tele...	MINNEAPOLIS (WCCO) — A Fridley man has been ch...	1
Finland jails police chief Aarnio for drug-smu...	Published\n\nimage copyrightAFP\n\nA Finnish c...	1
'Britain's sickest love rat' serial fraudster ...	A SERIAL fraudster once branded Britain's sick...	1
Kanye West's strange presidential bid unravels...	(CNN) Kanye West is on the ballot in Minnesota...	0
The man who would be king	God's Fury, England's Fire: A New History of t...	0

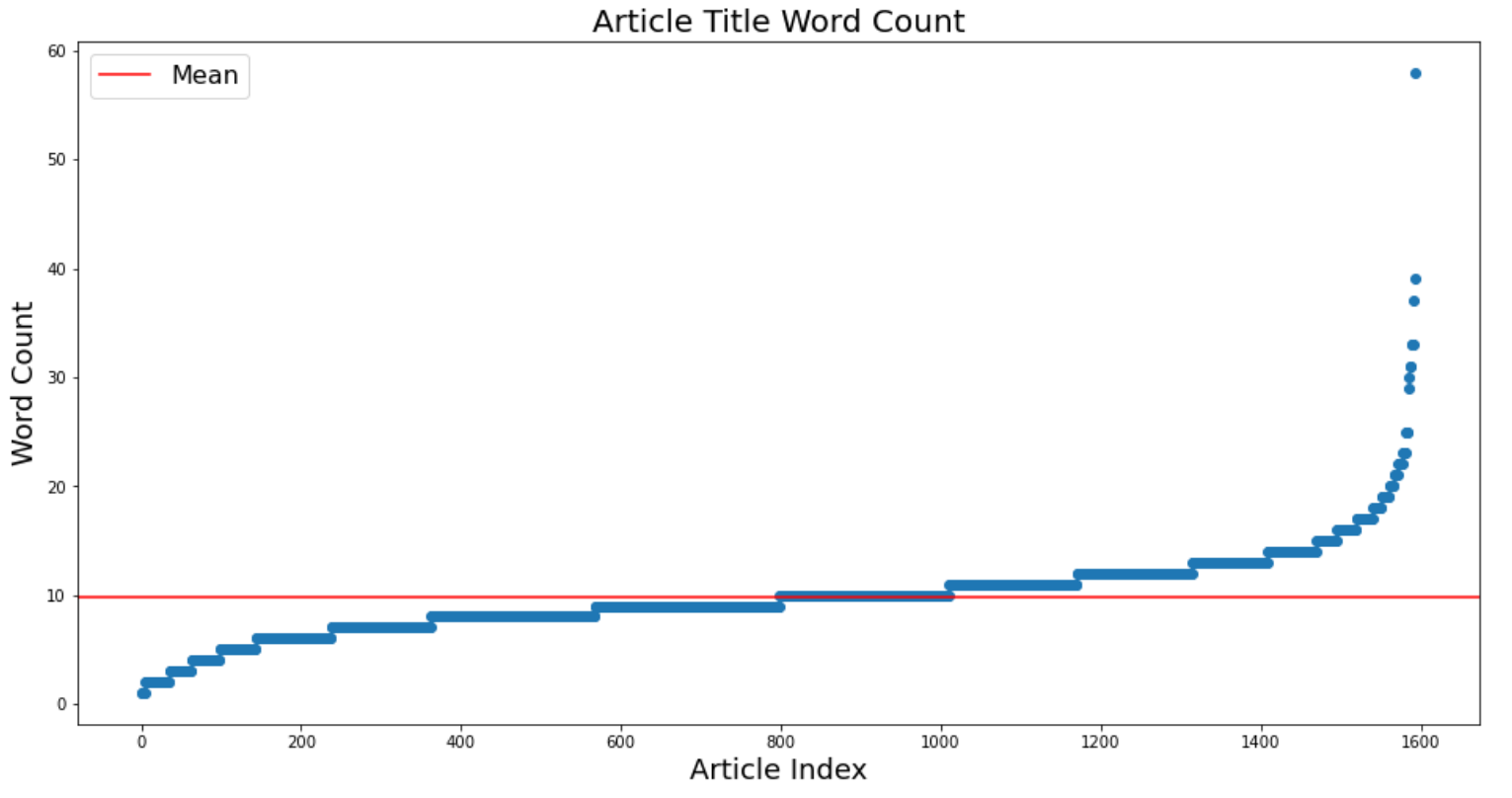
- For validation:
  - 97 AM
  - 62 NAM/random

- For final test result (client data):
  - Around 100 AM, 50 NAM, 50 random

Article Body Word Count

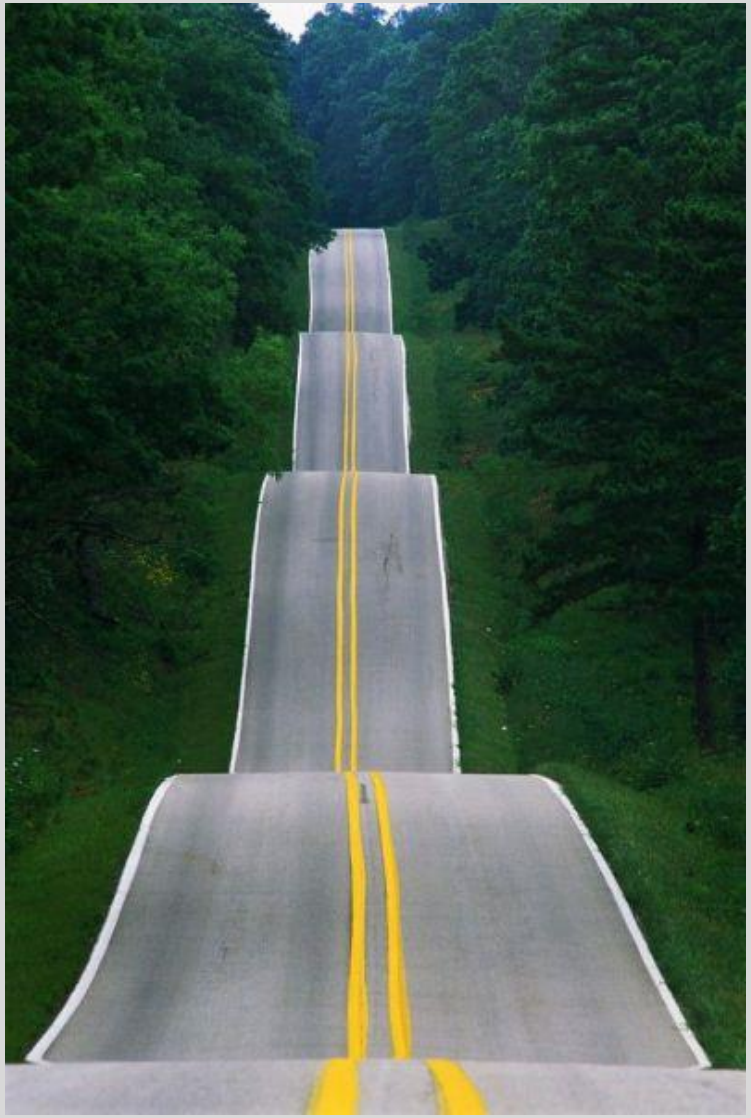


Article Title Word Count



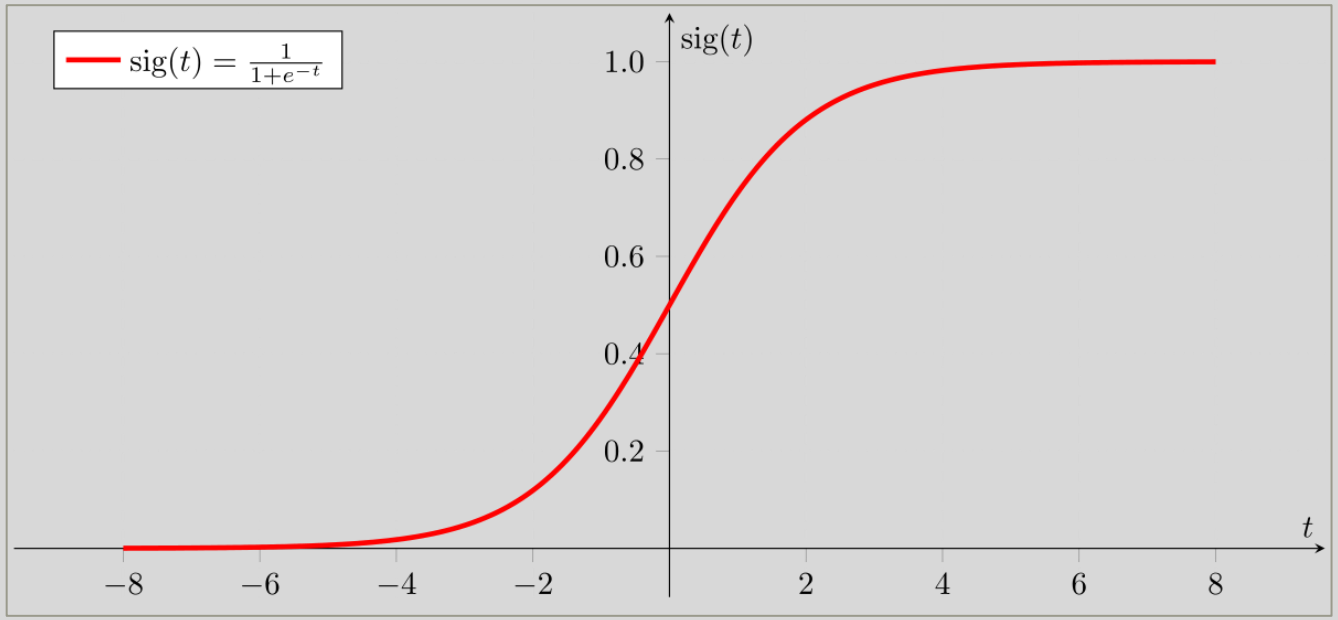
Model Path

- Naïve Bayes
- **Logistic regression**
  - Word vectorizers
  - Solvers
  - Ensembles
- BERT
- **RoBERTa**
  - Hard voting over windows
  - Soft voting over windows
- **Ensembles**
  - Logistic regression + RoBERTa-All
  - Logistic regression + RoBERTa-Title + RoBERTa-Body

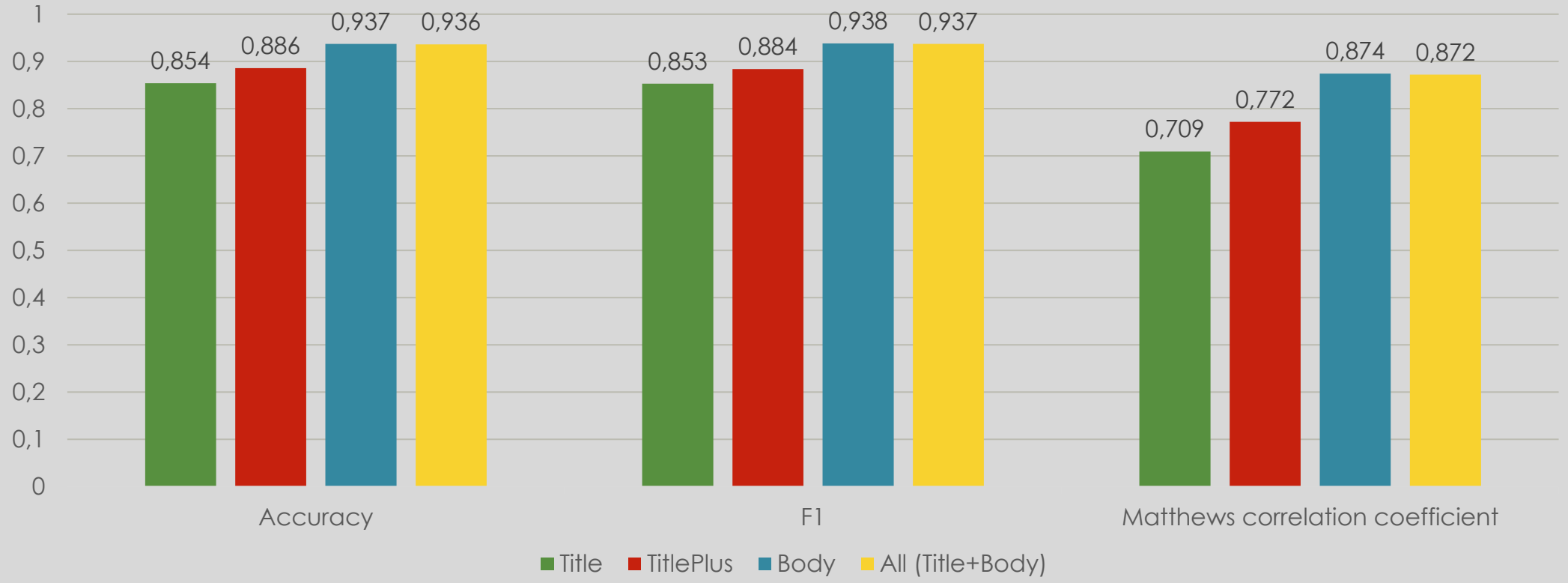


Logistic Regression (LogReg)

- Word Vectorizer
  - Doc2Vec
  - CountVectorizer
  - **TfidfVectorizer**
    - **n-gram = 1**
- Solver
  - Liblinear
  - Saga
  - **Lbfgs**
- Hyperparameter testing
  - Penalty = L2
  - C = 17 (regularization strength)
- Ensembles
  - BaggingClassifier
  - AdaBoostClassifier

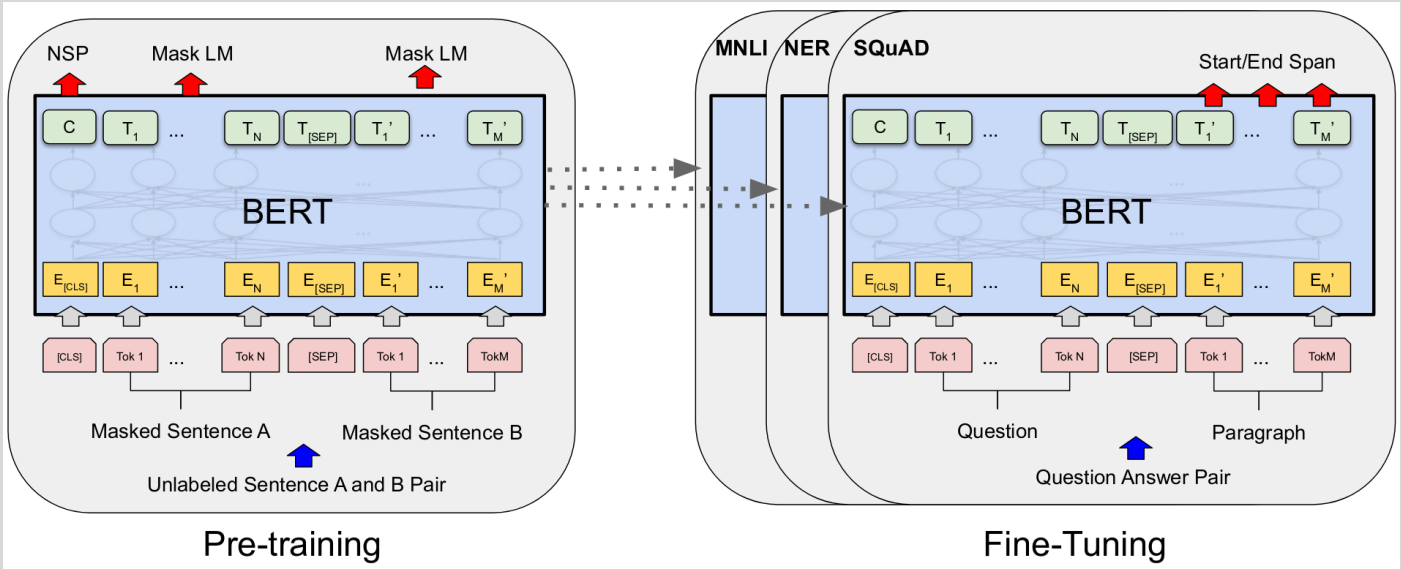


Logistic Regression (LogReg) - Results



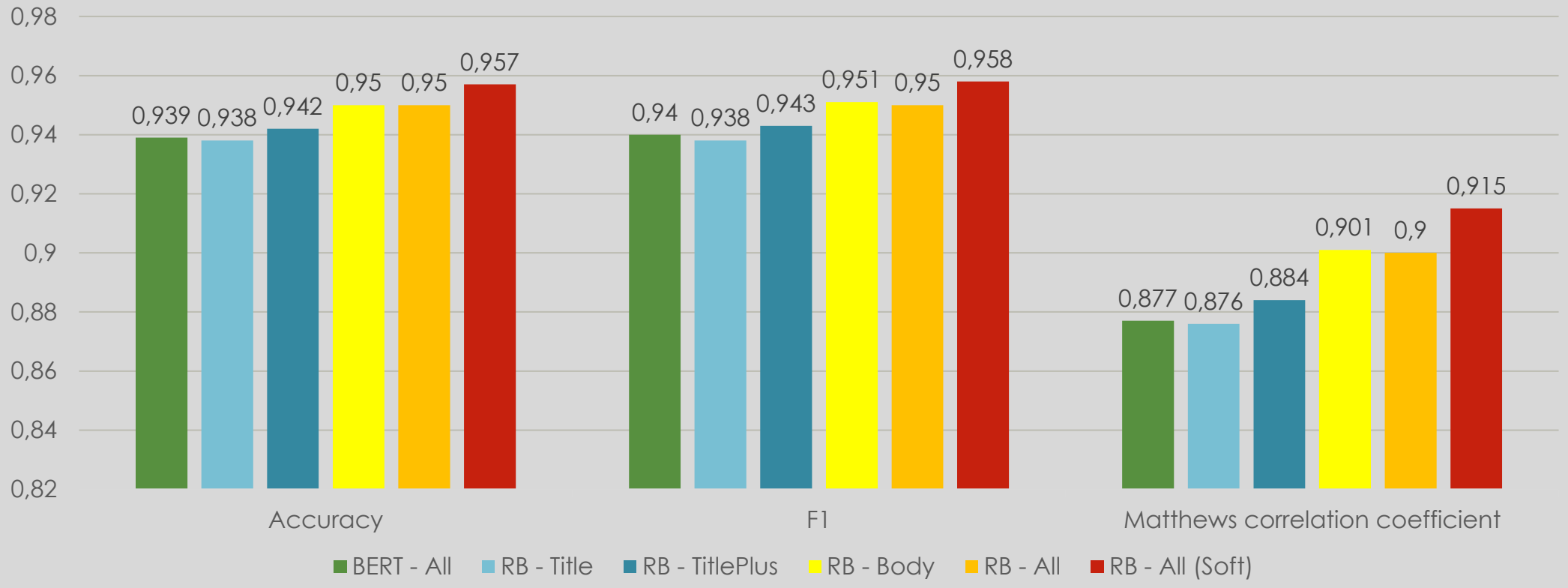
RoBERTa (RB)

- Improved BERT
- State of the art NLP deep learning model
  - Pre-trained on huge data set
- Requires only fine tuning
- Version: Base (125M params)
- Hyperparameters
  - Learning rate
  - Epoch
- Sliding windows (20% overlap)
  - Hard vote over windows
  - Soft vote over windows



Note: RoBERTa does not pre-train with next sentence prediction (NSP) as shown in picture (only BERT does)

RoBERTa (RB) and BERT

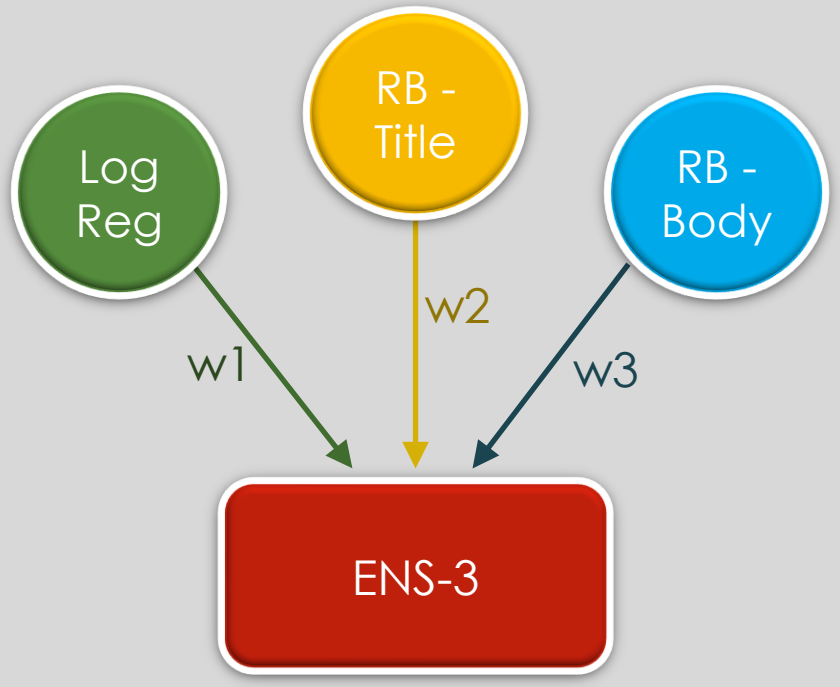


Ensembles (ENS)

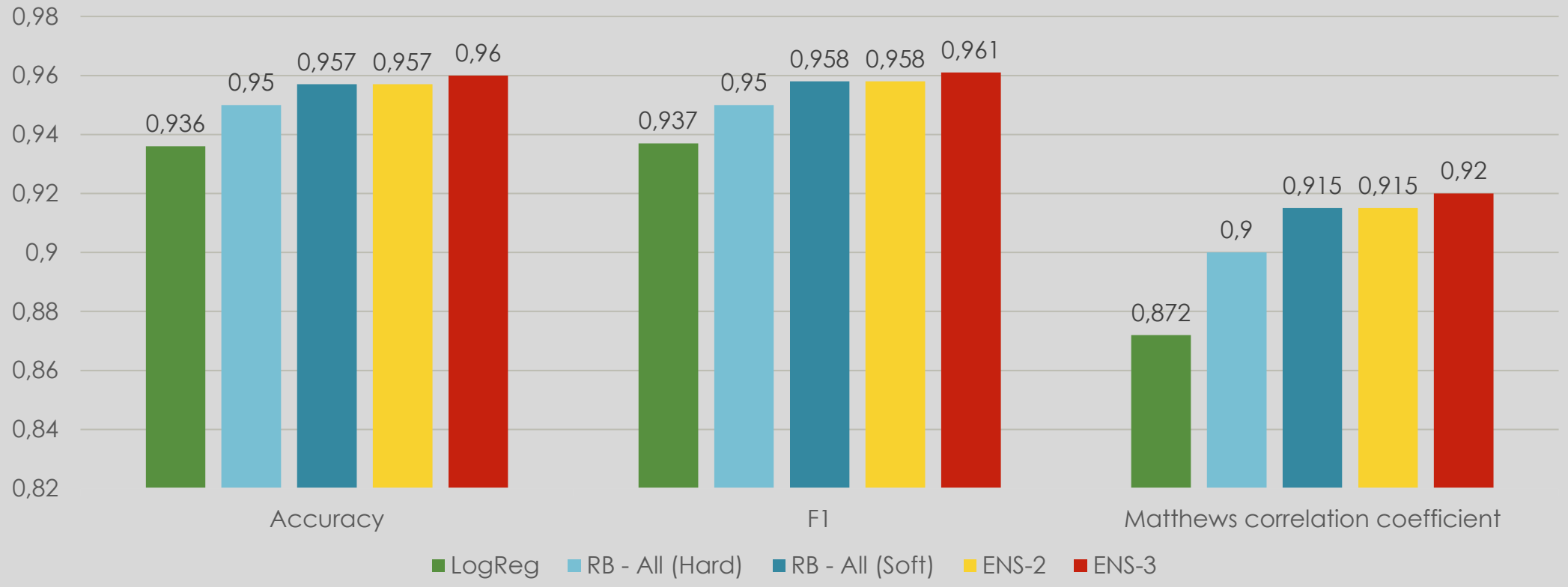
- Analysis showed models were not making same mistakes

- Ensemble-2 (**ENS-2**):
  - **Logistic regression – All** + **RoBERTa – All**
  - Soft Voting
    - Probabilities
    - Weighted (MCC scores)

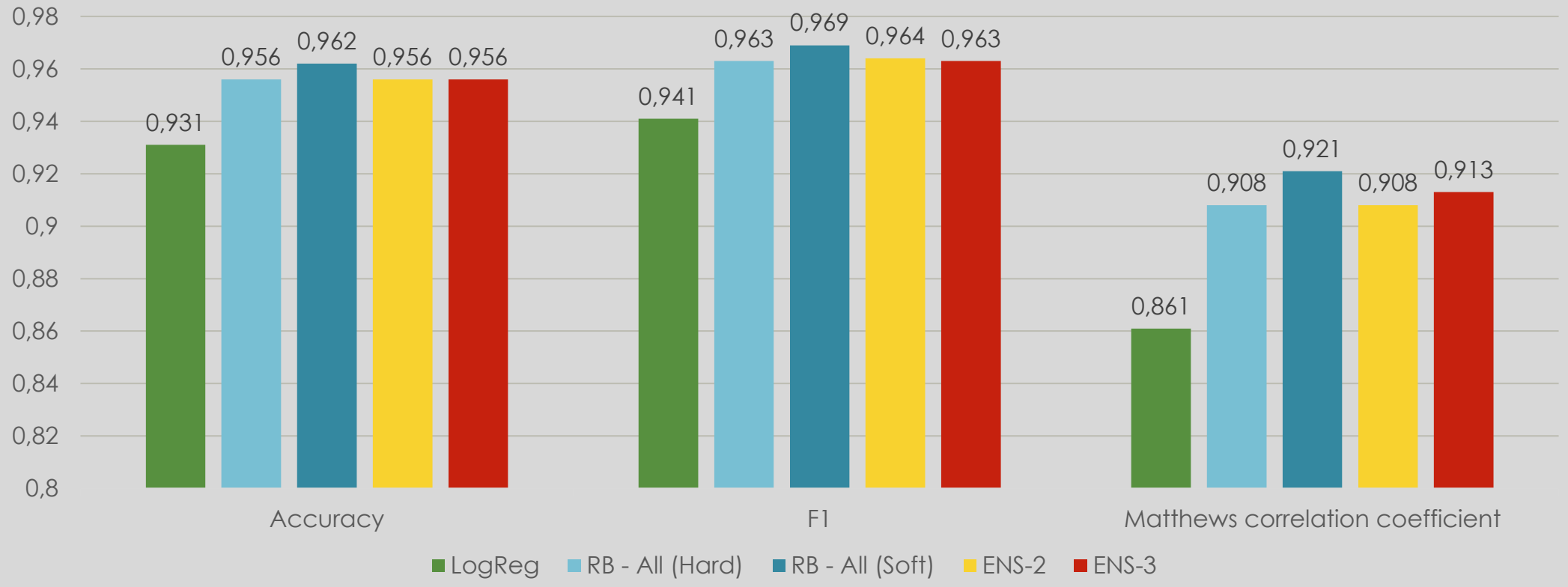
- Ensemble-3 (**ENS-3**):
  - **Logistic regression – All** + **RoBERTa – Title** + **RoBERTa – Body**
  - Soft Voting
    - Probabilities
    - Weighted (MCC scores)



Results – Cross Validation (6 Fold)



Results – Test data



Time Performance

	LogReg	RB - All (Hard)	RB - All (Soft)	ENS-2	ENS-3
Execution Time	68 ms	9.21 s	9.21 s	9.48 s	12.1 s
CV-6 Accuracy	93,6%	95%	95.7%	95.7%	96%

Run on test set (159 articles)

Executed in Google Colab

- Logistic regression is significantly faster
- RoBERTa (base) needs around 10GB of GPU memory

Ways to improve

- More data better results
  - When gathering additional data we saw improvements for Logistic regression and RoBERTa
- Bigger batch size (less overfitting)
- RoBERTa large (requires a lot of GPU memory)

Github Repository

https://github.com/dannoc96/4N-Negative-News-Neural-Nets---ML-Project