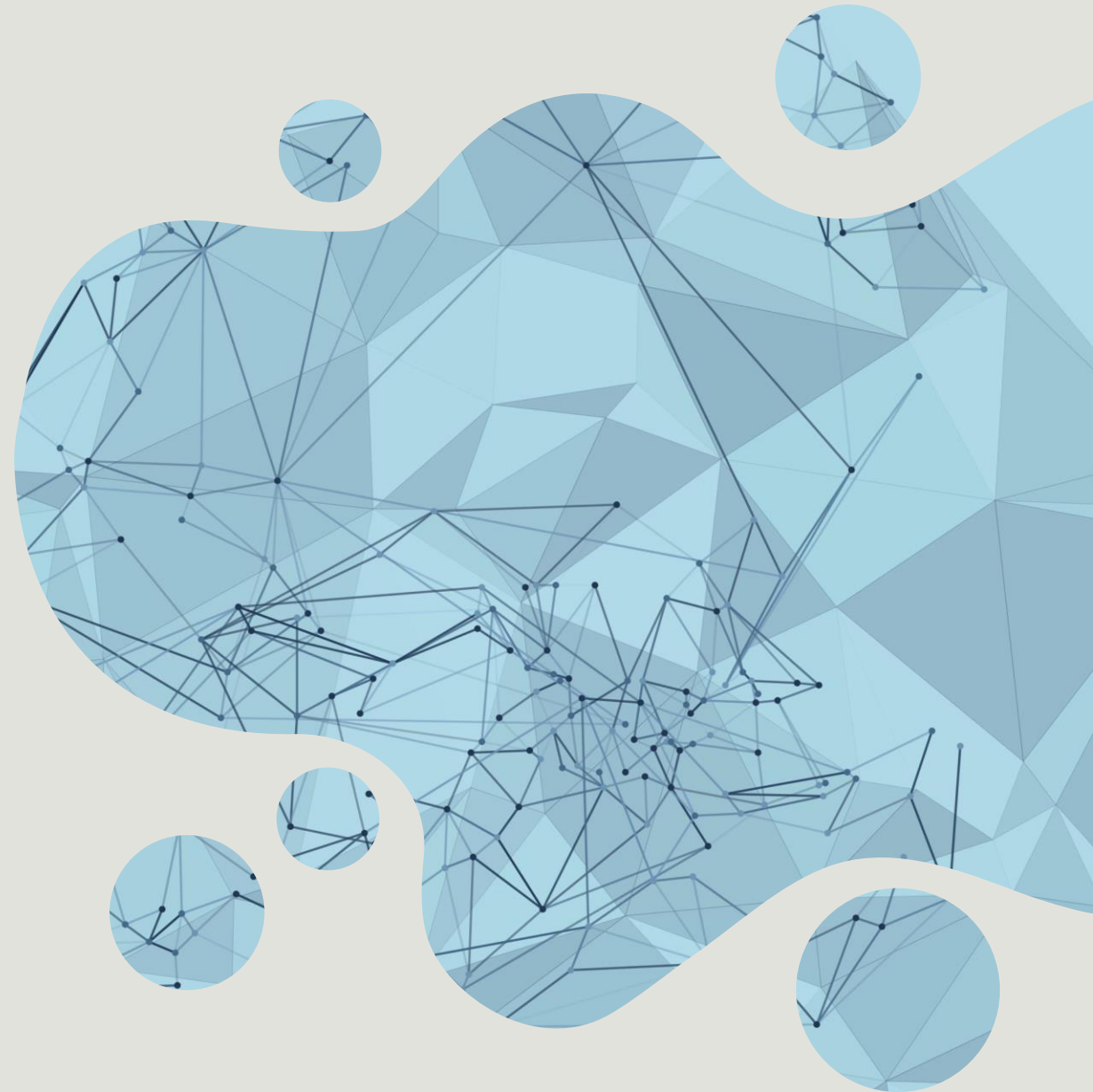


Anomaly detection on Scientific Publications

David-Gabriel ION

Thesis advisor:

Prof. dr. ing. Ciprian DOBRE



Context

The problem

Determine if an author wrote a given scientific publication

Intuition

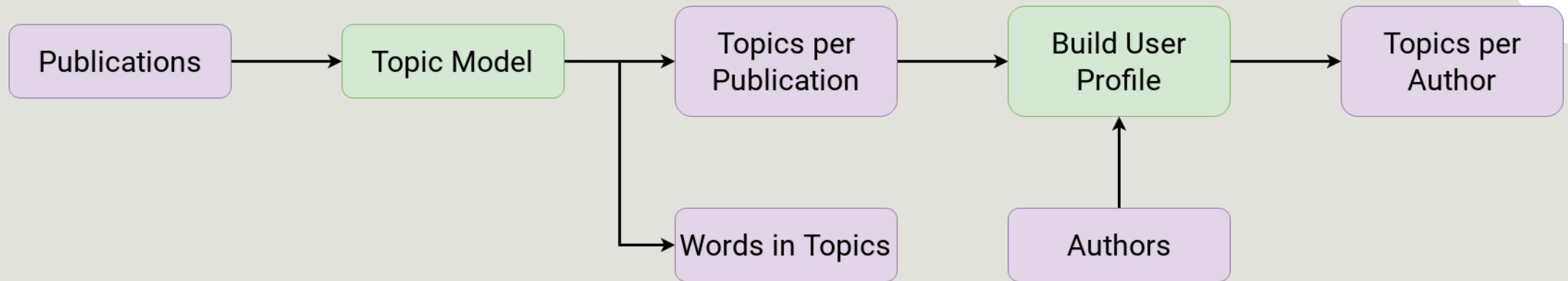
An author only writes publications in their area of expertise

General solution

Determine what are the areas of expertise for the given authors, given their past published papers

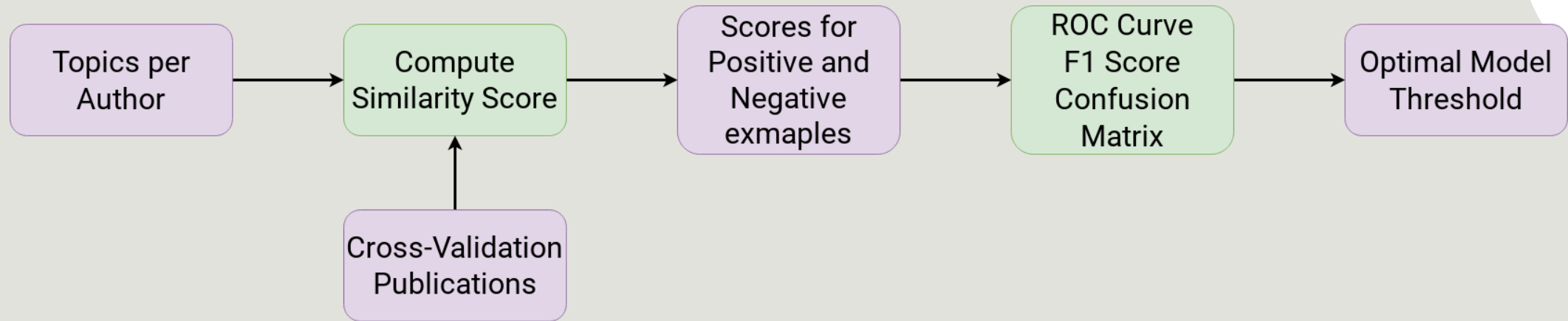
Check if a new publication is about a field unrelated to a given author

Training



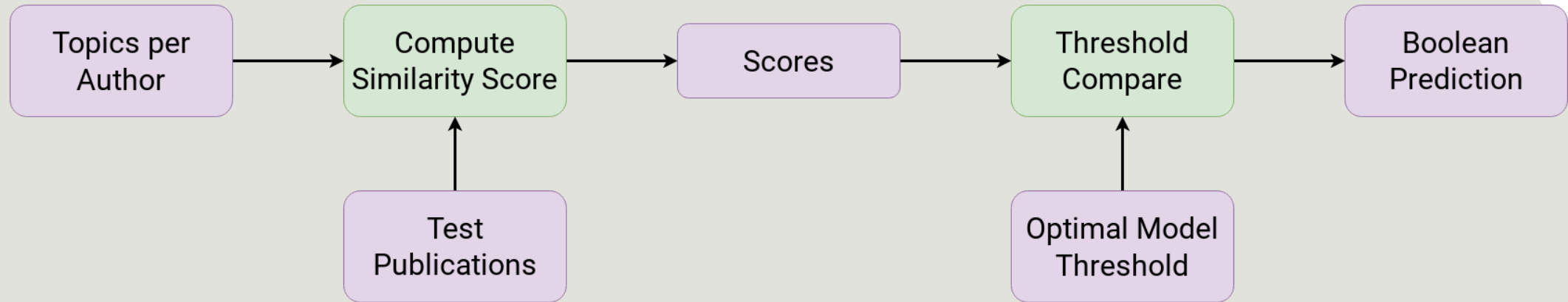
- Topic Model is Latent Dirichlet Allocation (will be replaced by BERT)
- User profile is the mean and standard deviation of the authored publications

Tuning



- Similarity Score is the Normal Probability Density Function

Testing



Discovered topics

paper algorithm
result base
system
design gener
model simul
propos perform problem
control order
method time
function two optim

propos paper
solut perform user
data imag
provid base model
network
inform comput
system
implement present
servic architectur

reaction experiment
magnet
obtain liquid determin
process
studi phase
water
acid extract result
concentr
temperatur field

investig microscopi
property
film structur
composit
result
surfac alloy
character deposit sampl
obtain oxid
metal studi

present oper fuel
air current high
energi
paper gener heat
perform plant system measur
power consumpt
electr increas
result

numer studi
structur surfac stress
model method
simul mechan obtain
flow paper forc
present analysi
result two
experiment determin

paper learn
one social technolog
develop
also activ risk
research
new knowledg manag
present organ human
student
educ studi

studi treatment
develop result food
CELL product new
technolog applic. activ
device high materi
process structur current

manufactur robot
present
research
product
materi machin part technolog
process industri
test result
paper develop
system design
univers control

particl
model result
observ data
rang mass detector two
measur plasma
energi optic
state field
decay fusion highcollis

Discovered topics

paper algorithm
result base
system
design gener
model simul
propos perform problem
control time
method optim
function two

propos paper
solut perform user
data imag
provid base model
network
inform comput
system
implement present
servic architectur

reaction experiment
magnet
obtain liquid determin
process
studi phase
water
acid extract result
concentr
temperatur field

investig microscopi
property
film structur
composit
result
surfac deposit
obtain electron sampl
character oxid
metal studi

present oper fuel
air current high
energi
paper gener heat
perform plant system measur
power consumpt increas
electr result

numer studi
structur surfac stress
model method
simul mechan obtain
flow paper forc
present analysi
result two
experiment determin

paper learn
one social technolog
develop
also activ risk
research
new knowledg manag
present organ human
student
educ studi

studi treatment
develop result food
CELL product present new
technolog applic. activ
device high materi
process structur current

manufactur robot
present
research
product
materi machin part technolog
process result industri
paper work develop
system univers control design

particl
model result
observ data
rang mass detector two
measur plasma
energi optic
state field
decay highcollis

Discovered topics

paper algorithm
result base
system
design gener
model simul
propos perform problem
control time
method optim
function two

propos paper
solut perform user
data imag
provid base model
network
inform comput
system
implement present
servic architectur

reaction experiment
magnet
obtain liquid determin
process
studi phase model
acid extract water
result concentr
temperatur field

investig microscopi
film structur
composit
result alloy
surfac deposit sampl
electron oxid
obtain
materi metal studi

present oper fuel
air current high
energi
paper gener heat
perform plant system measur
power
effici consumpt increas
electr result

numer studi
structur surfac stress
element
model
simul mechan method
flow obtain
present paper forc
result analysi two
experiment determin

paper learn
one social technolog
develop
also activ risk
research
new knowledg manag
present organ human
student
educ studi

studi treatment
develop result food
wast new
product present
CELL
applic. activ
process materi
structur current

manufactur robot
present
materi research
machin part technolog
product
test result industri
process
work develop
paper
system design
univers control

particl
model result data
observ event
rang mass detector two
measur plasma
decay fusion
energi optic
state field
highcollis

Discovered topics

paper algorithm
result base
system
design gener
model simul
propos perform problem order
control
method time
function two optim

propos paper
solut perform user
data imag
provid base model
network
inform comput
system
implement present
servic architectur

reaction experiment
magnet
obtain liquid determin
process
studi phase model
acid extract water
result concentr
temperatur field

investig microscopi
property
film structur
composit
result surfac alloy
character deposit sampl
obtain
material oxid metal studi

present oper fuel
air current high
energi
paper gener heat
perform system measur
plant power
effici consumpt increas
electr result

numer studi
structur surfac stress
element model
simul mechan method
flow
paper obtain
present result two
experiment

paper learn
one social technolog
develop
also activ risk
research
new knowledg manag
present organ human
import student
educ studi

studi treatment
develop result food
wast present new
product
CELL
applic. device
process materi
high effect

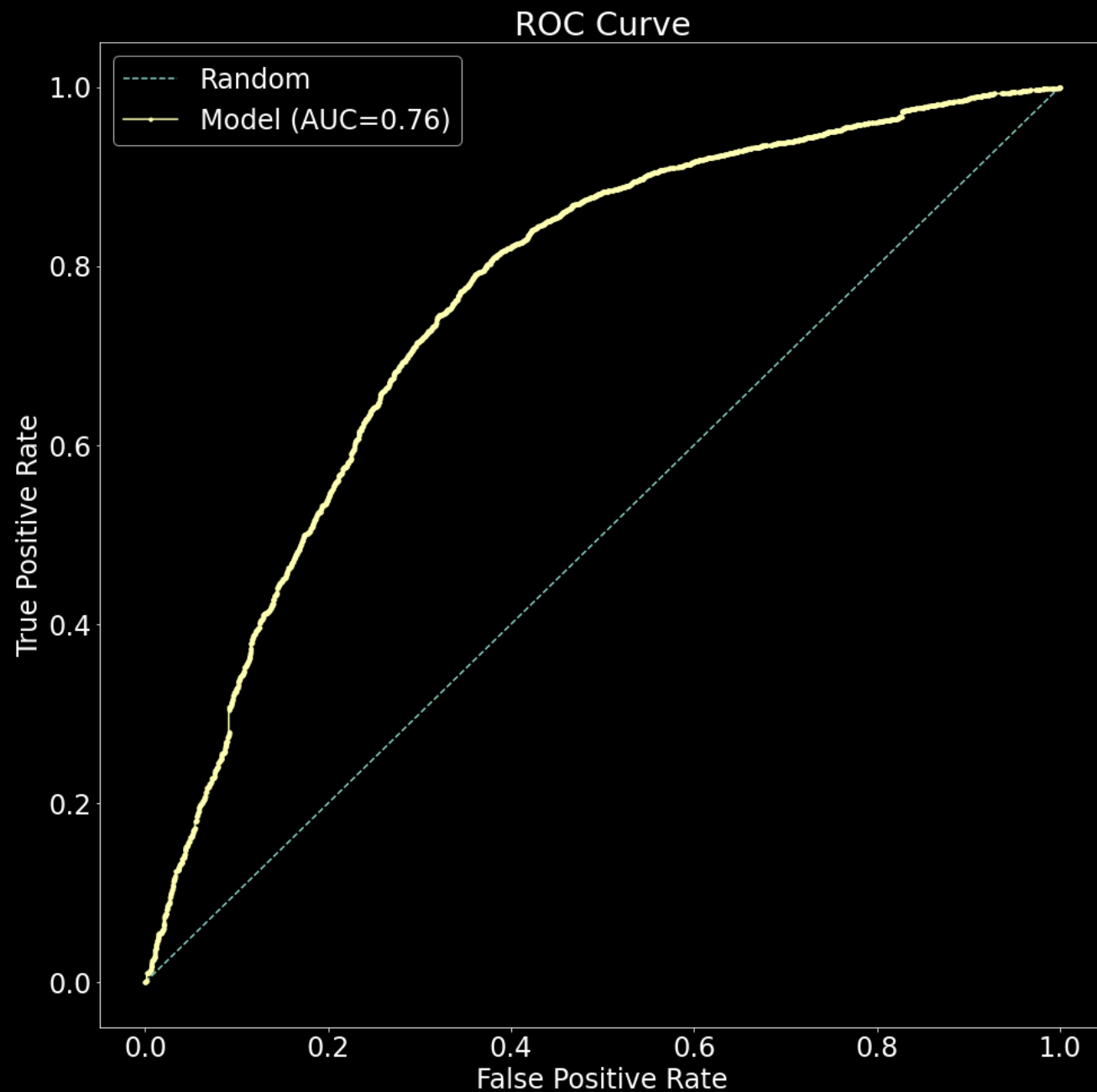
manufactur robot
present
research
materi machin
part technolog
product
test result industri
process
work develop
paper
univers control
system design

particl
model result data
observ
rang mass detector two
measur
decay fusion
energi
state optic
highcollis field

ROC Curve

Using the real negative examples

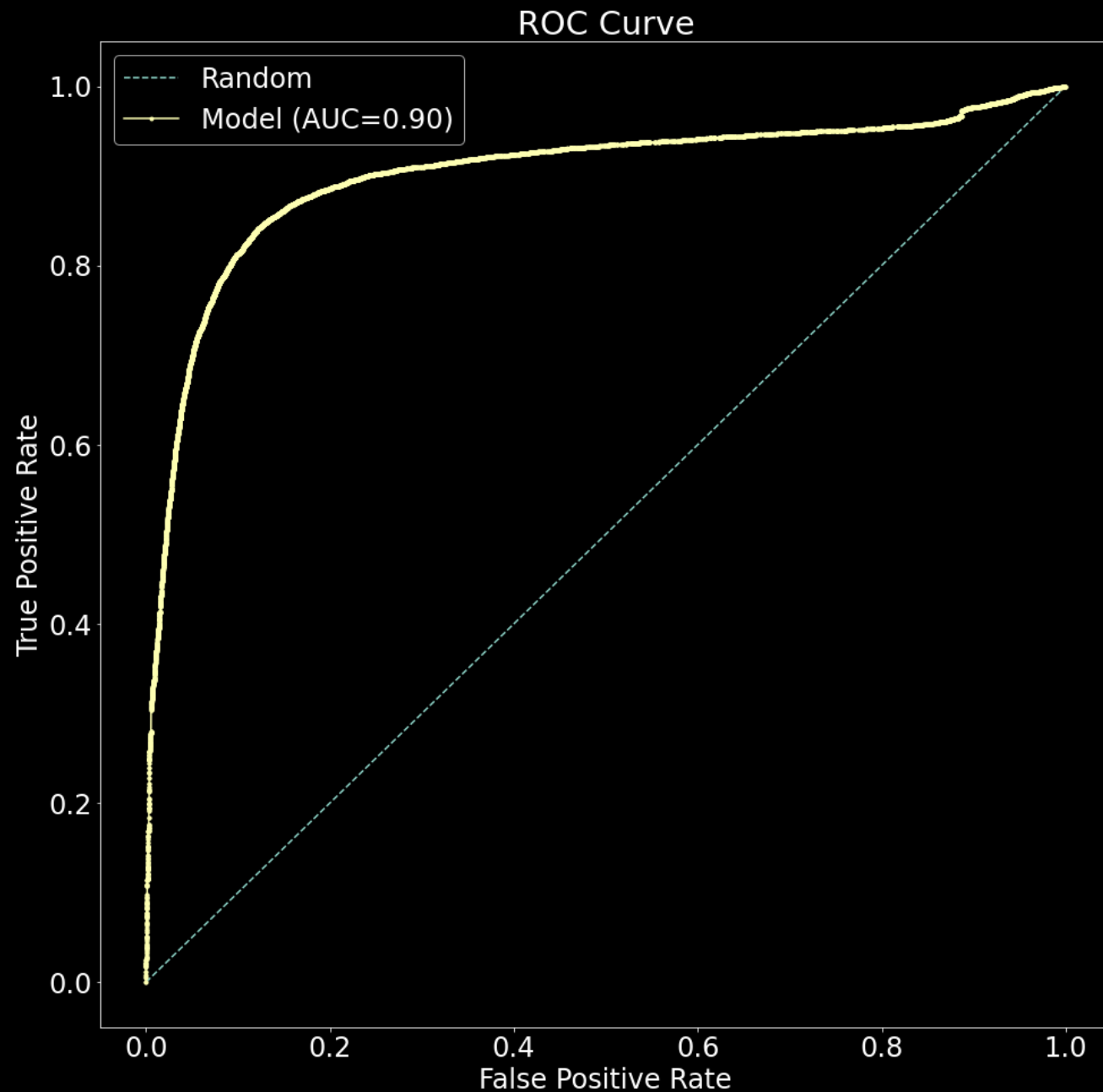
- Imbalanced dataset (7:1 ratio)
- Harder to tune
- Could have used FBeta score, but that introduces another parameter



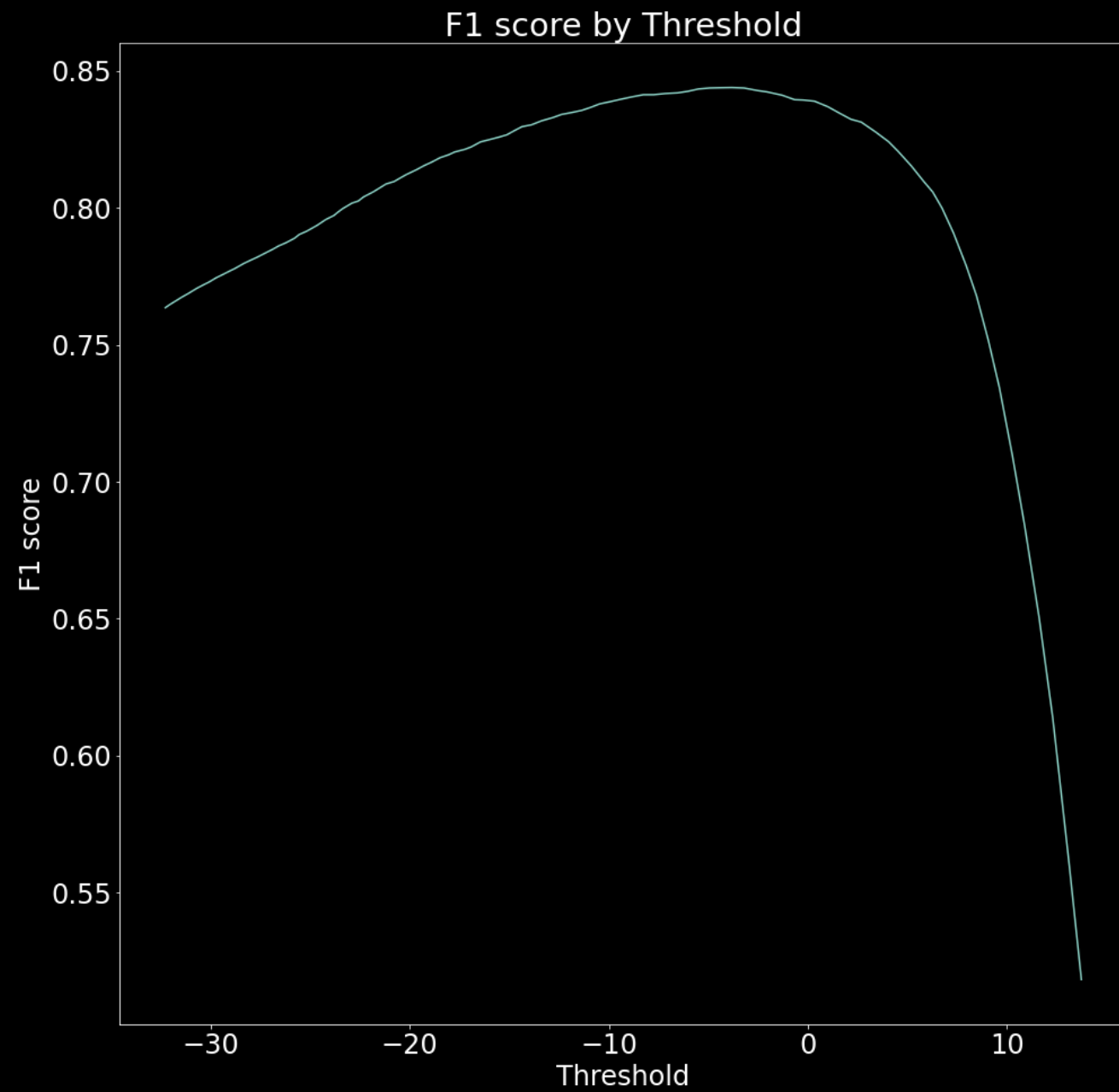
ROC Curve

Using random publication-author pairs

- Easier to tune using F1 score



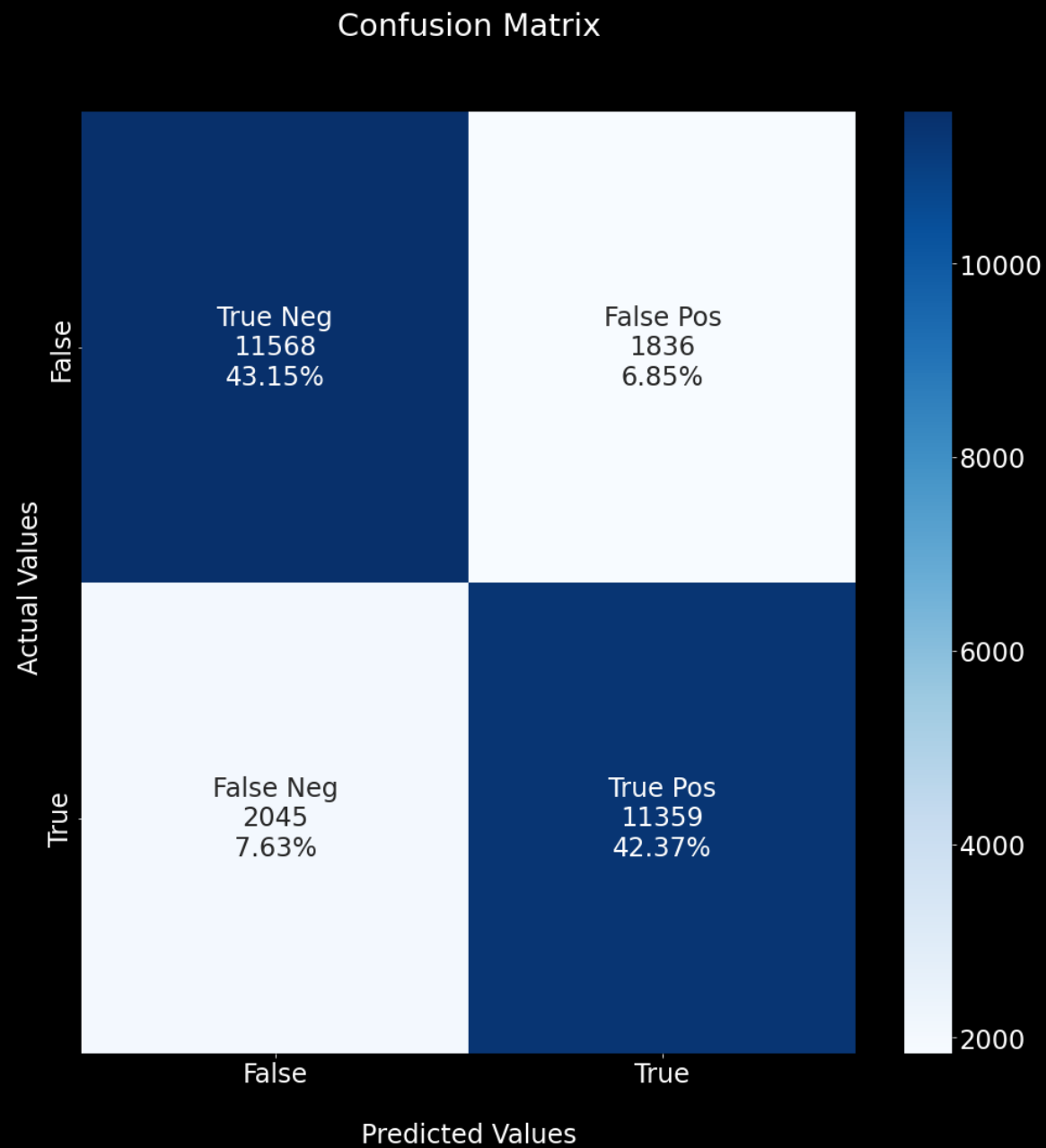
F1 Score



Confusion Matrix

On the synthetic dataset

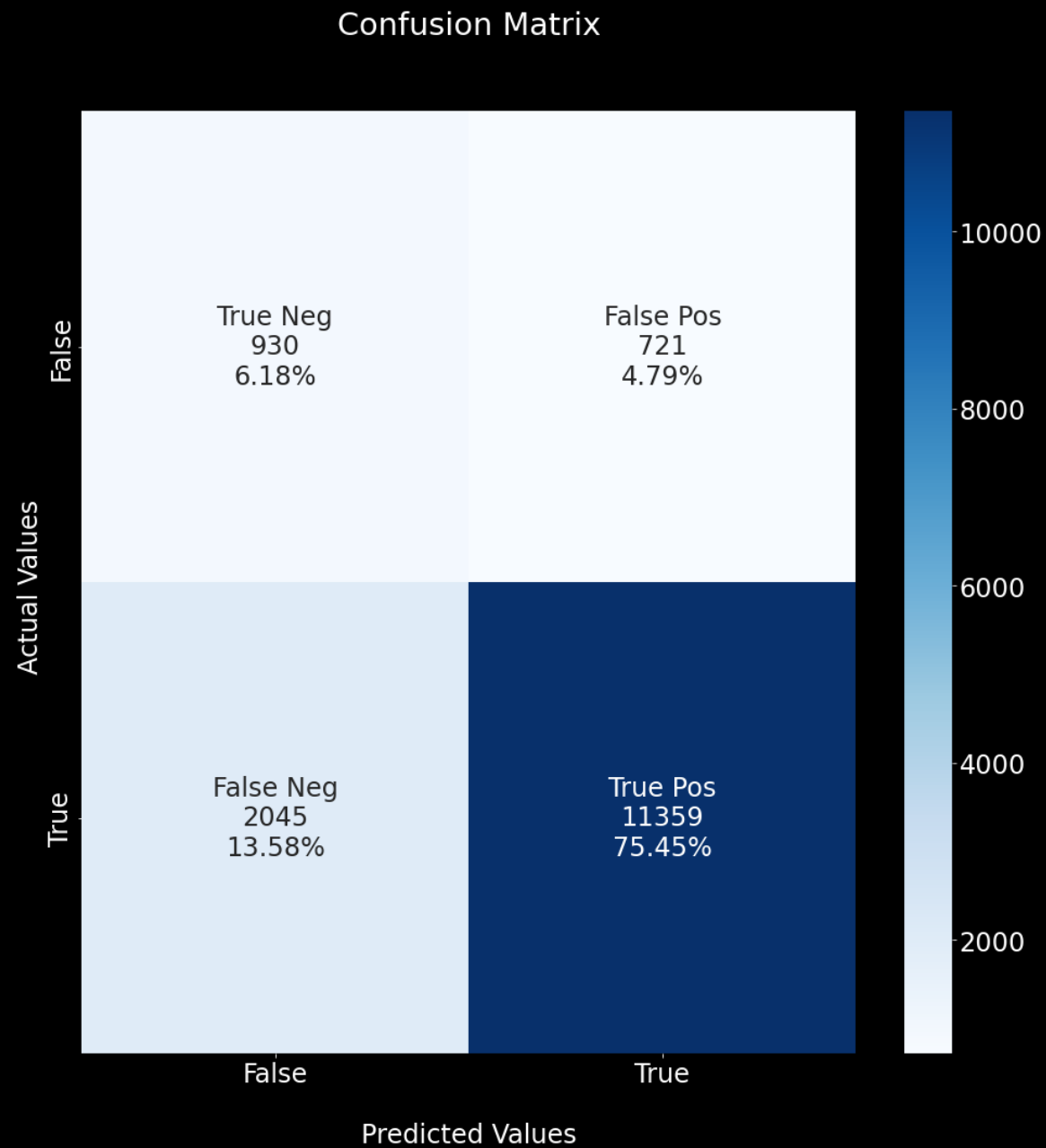
- 86% accuracy



Confusion Matrix

On the cross-validation dataset

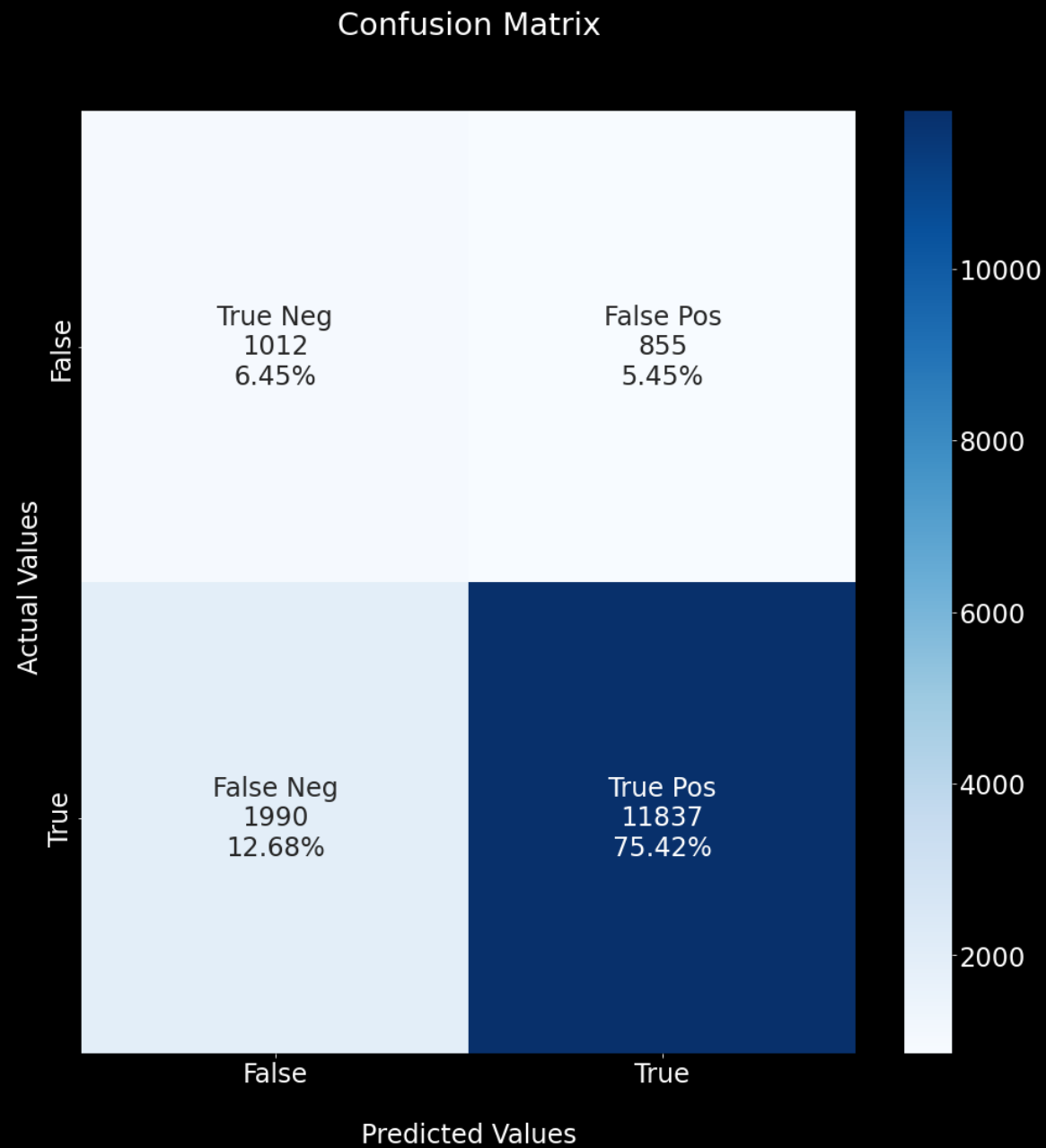
- 81% accuracy
- 0.84 F1 score



Confusion Matrix

On the test dataset

- 81% accuracy
- 0.89 F1 score



Conclusions

Limitations

- Word Ambiguity
- Small Dataset
- No end-to-end training

Improvement plans

- Replace LDA with BERT
- Implement end-to-end supervised training



Thank you!