Fake News Detector

Eithar Abdalslam

Manuel Cecere

Antoine Luu

Anastasia Cotov

01

Goals

BUILD A FAKE NEWS DETECTOR ON TWITTER

ANALYZE THE MOST USEFUL FEATURES

DEPLOY OUR MODEL ON REAL-WORLD DATA

Our Beatiful Team

ANTOINE

ML ENGINEER

MANUEL

GLOVE MODEL

EITHAR

DATA ANALYSIS

ANASTASIA

LASER MODEL

03

Pheme Dataset

Events	Threads	Tweets	Rumours	Non-rumours	True	False	Unverified
Charlie Hebdo	2,079	38,268	458	1,621	193	116	149
Sydney siege	1,221	23,996	522	699	382	86	54
Ferguson	1,143	24,175	284	859	10	8	266
Ottawa shooting	890	12,284	470	420	329	72	69
Germanwings-crash	469	4,489	238	231	94	111	33
Putin missing	238	835	126	112	0	9	117
Prince Toronto	233	902	229	4	0	222	7
Gurlitt	138	179	61	77	59	0	2
Ebola Essien	14	226	14	0	0	14	0
Total	6,425	105,354	2,402	4,023	1,067	638	697

"All-in-one: Multi-task Learning for Rumour Stance classification,Detection and Verification" by E. Kochkina, M. Liakata, A. Zubiaga

Pheme Dataset

Events	Threads	Tweets	Rumours	Non-rumours	True	False	Unverified
Charlie Hebdo	2,079	38,268	458	1,621	193	116	149
Sydney siege	1,221	23,996	522	699	382	86	54
Ferguson	1,143	24,175	284	859	10	8	266
Ottawa shooting	890	12,284	470	420	329	72	69
Germanwings-crash	469	4,489	238	231	94	111	33
Putin missing	238	835	126	112	0	9	117
Prince Toronto	233	902	229	4	0	222	7
Gurlitt	138	179	61	77	59	0	2
Ebola Essien	14	226	14	0	0	14	0
Total	6,425	105,354	2,402	4,023	1,067	638	697

Features available on the twitter api

+50

Fake News Detector

Multidisciplinary Project

Selected features

1 TEXT

2 LIKES COUNT

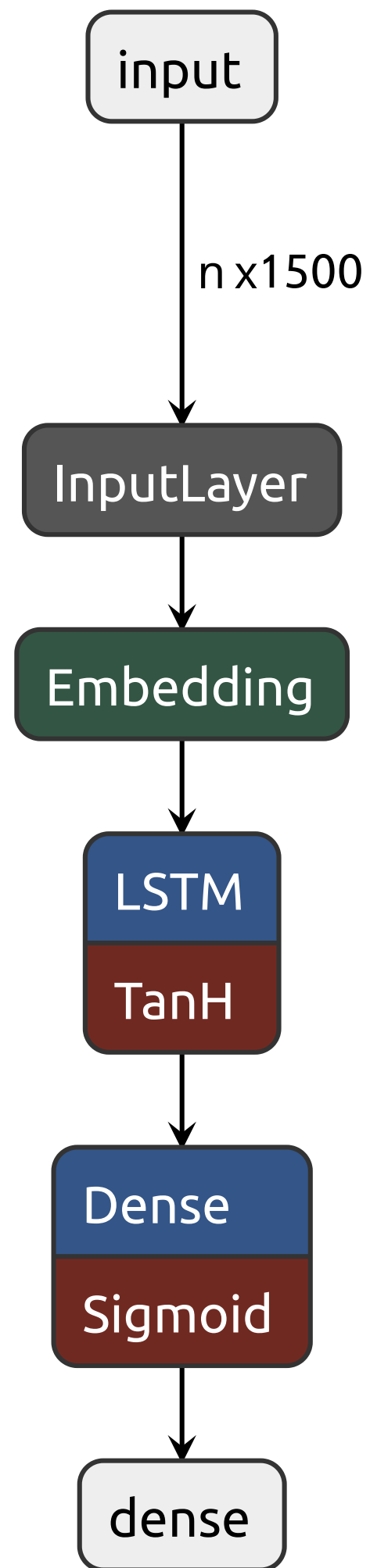3 RETWEETS

4 FOLLOWERS

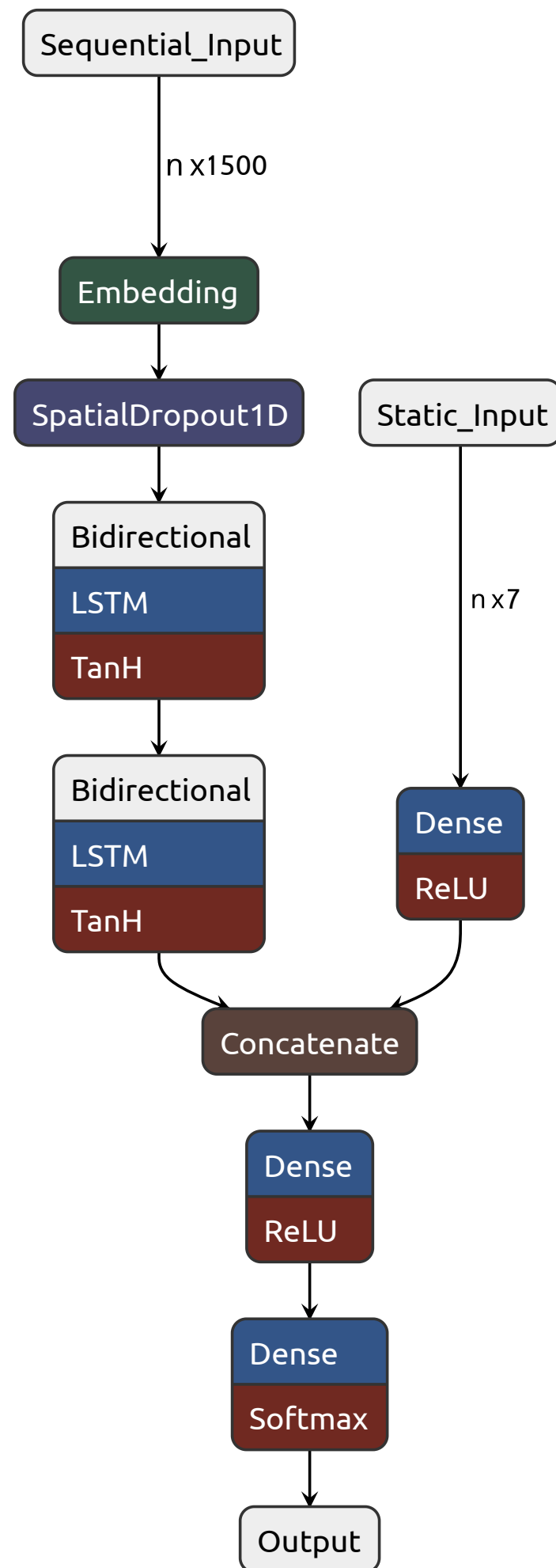5 FOLLOWINGS

6 TWEET COUNT

7 VERIFIED

8 Nº HASHTAGS
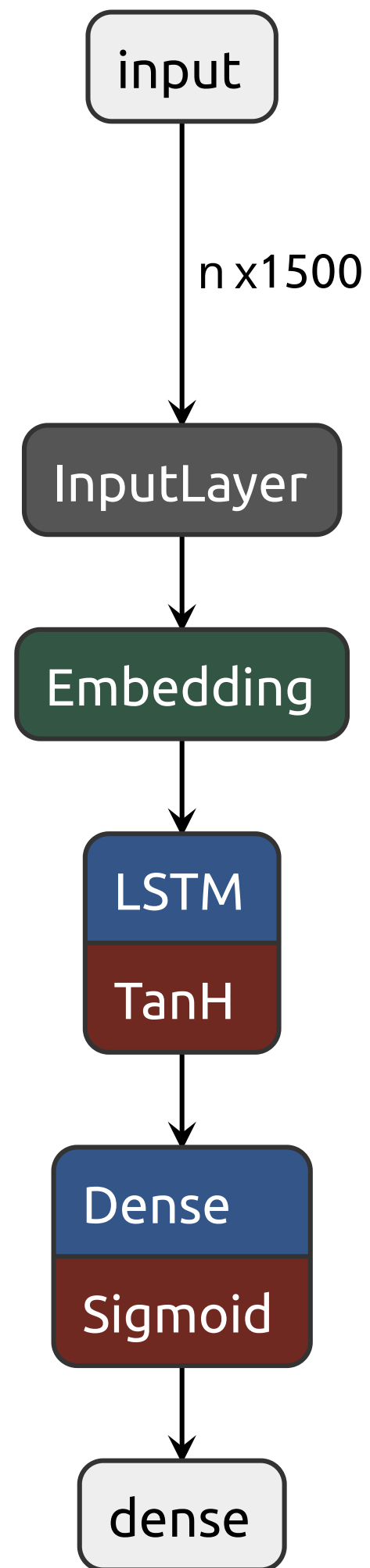
06

input

n x1500

InputLayer

Embedding

LSTM
TanH

Dense
Sigmoid

dense

Architectures used

07

input

n x1500

InputLayer

Embedding

LSTM
TanH

Dense
Sigmoid

dense

How about numerical features?

07

Sequential_Input

n x1500

Embedding

SpatialDropout1D

Static_Input

Bidirectional
LSTM
TanH

n x7

Bidirectional
LSTM
TanH

Dense
ReLU

Concatenate

Dense
ReLU

Dense
Softmax

Output

How about numerical features?

07

text

Embedding

embedded output

How about embedding?

english tweets

Glove

embedded outputs

How about embedding?

english tweets

↓

Glove

↓

embedded outputs
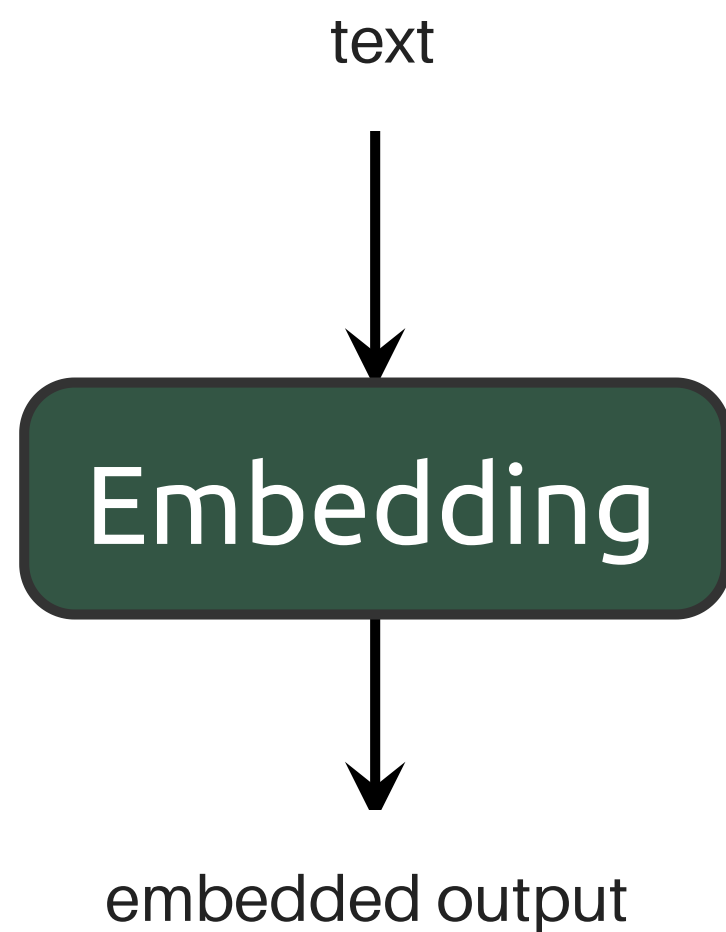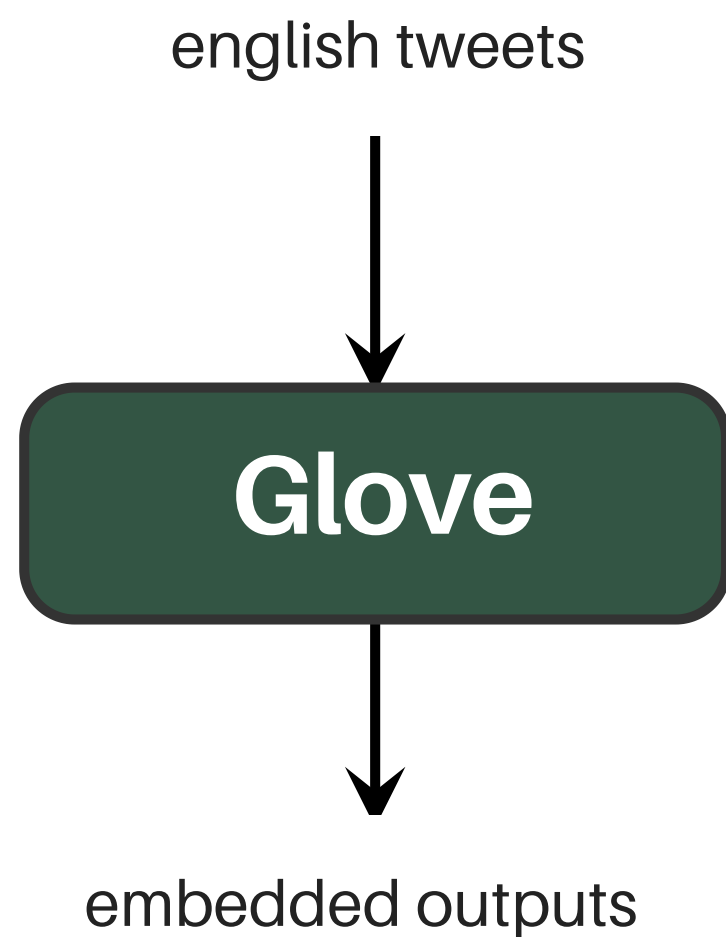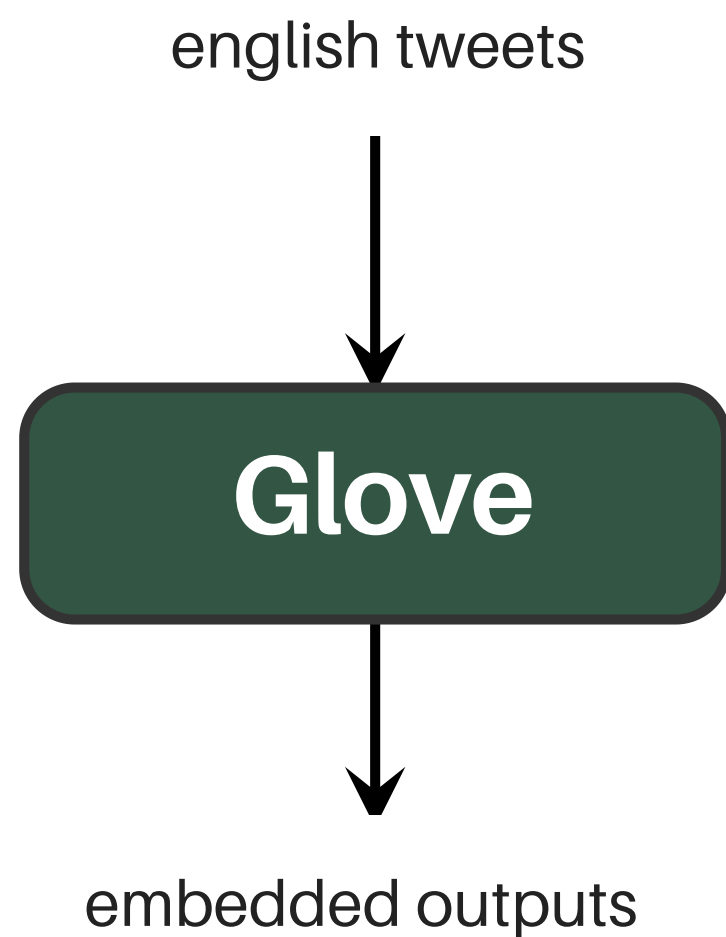
Can we manage more languages?

tweets

LASER

embedded outputs
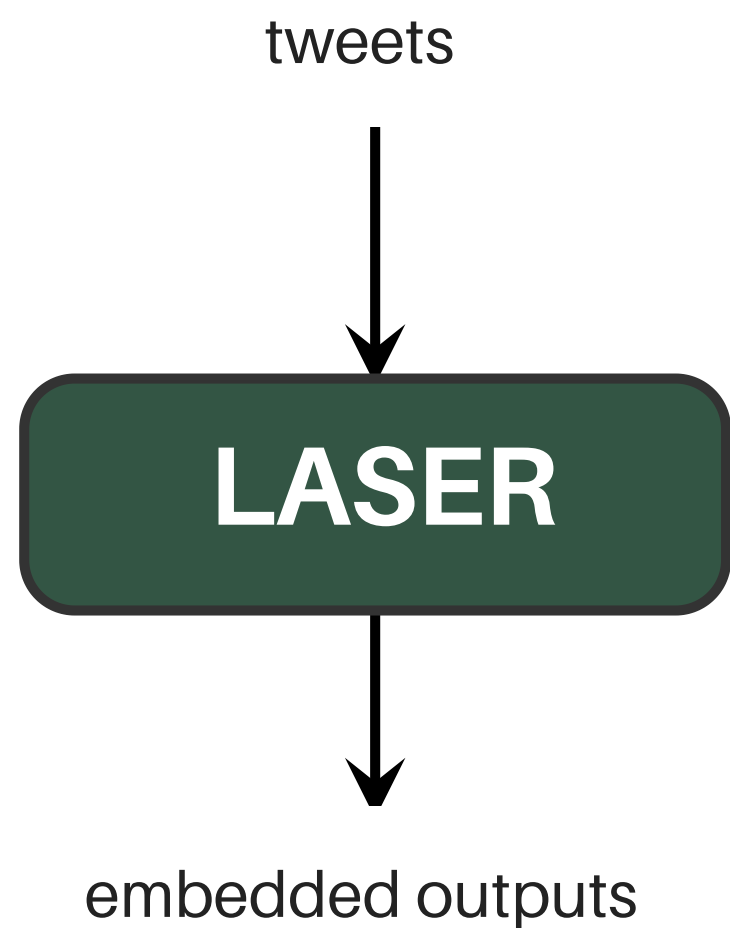
Language Agnostic Embedding

08

ACCURACY WITH RANDOM TRAIN-TEST SPLIT

09

Fake News Detector

Multidisciplinary Project

82% **ACCURACY WITH RANDOM TRAIN-TEST SPLIT**

Fake News Detector

Multidisciplinary Project

82% **ACCURACY WITH RANDOM TRAIN-TEST SPLIT**

09

Fake News Detector

Multidisciplinary Project

LEAVE ONE EVENT OUT CROSS VALIDATION

Charlie Hebdo	Sydney siege	Ottawa	Ebola Essien	Gurlitt
Charlie Hebdo	Sydney siege	Ottawa	Ebola Essien	Gurlitt
Charlie Hebdo	Sydney siege	Ottawa	Ebola Essien	Gurlitt

09

Fake News Detector

Multidisciplinary Project

82% **ACCURACY WITH RANDOM TRAIN-TEST SPLIT**

ACCURACY WITH LEAVE ONE EVENT OUT

09

Fake News Detector

Multidisciplinary Project

82% ACCURACY WITH RANDOM TRAIN-TEST SPLIT

37% ACCURACY WITH LEAVE ONE EVENT OUT

09

Fake News Detector

Multidisciplinary Project
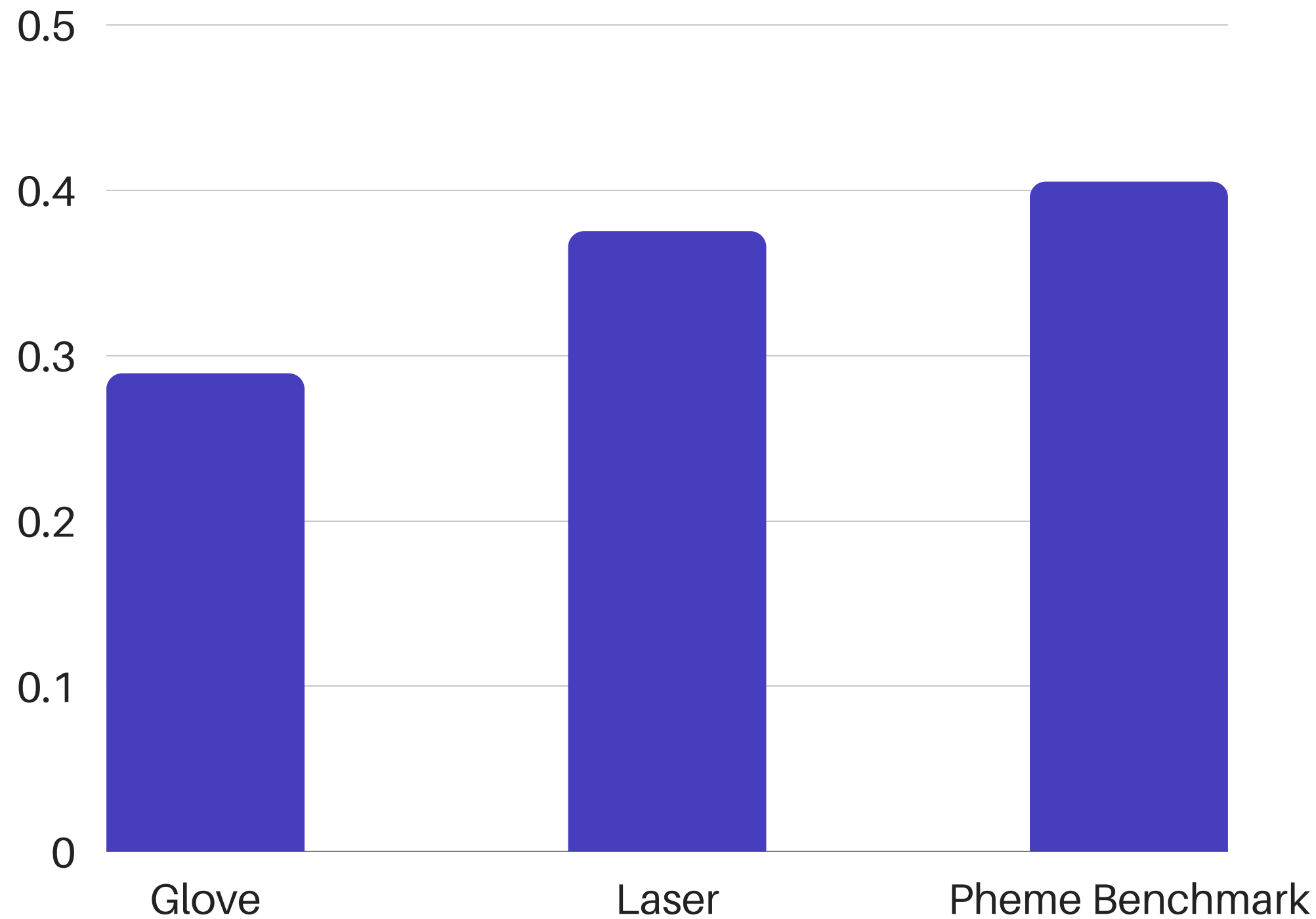
Accuracy per model

Accuracy using Leave One Event Out as Cross Validation.
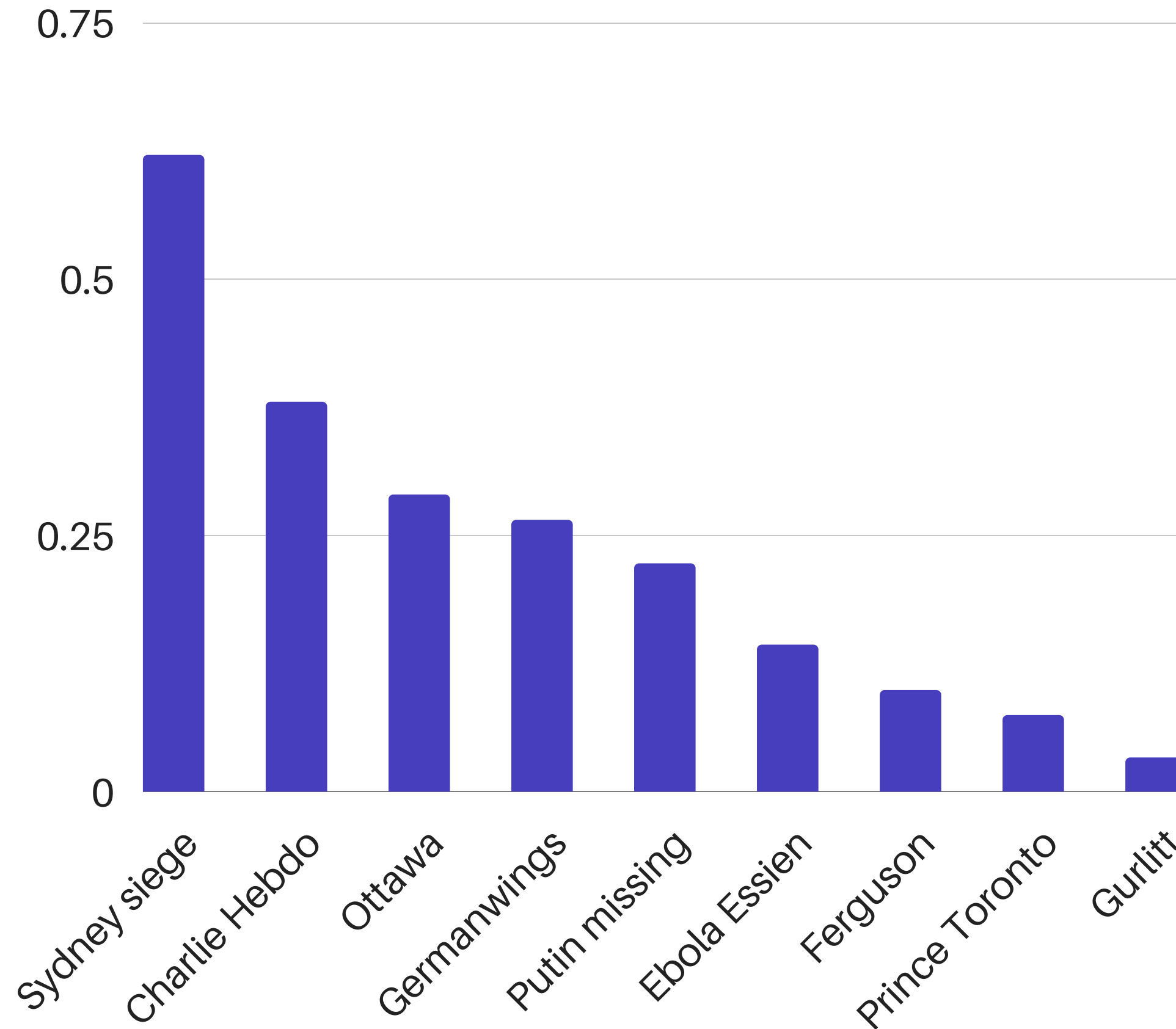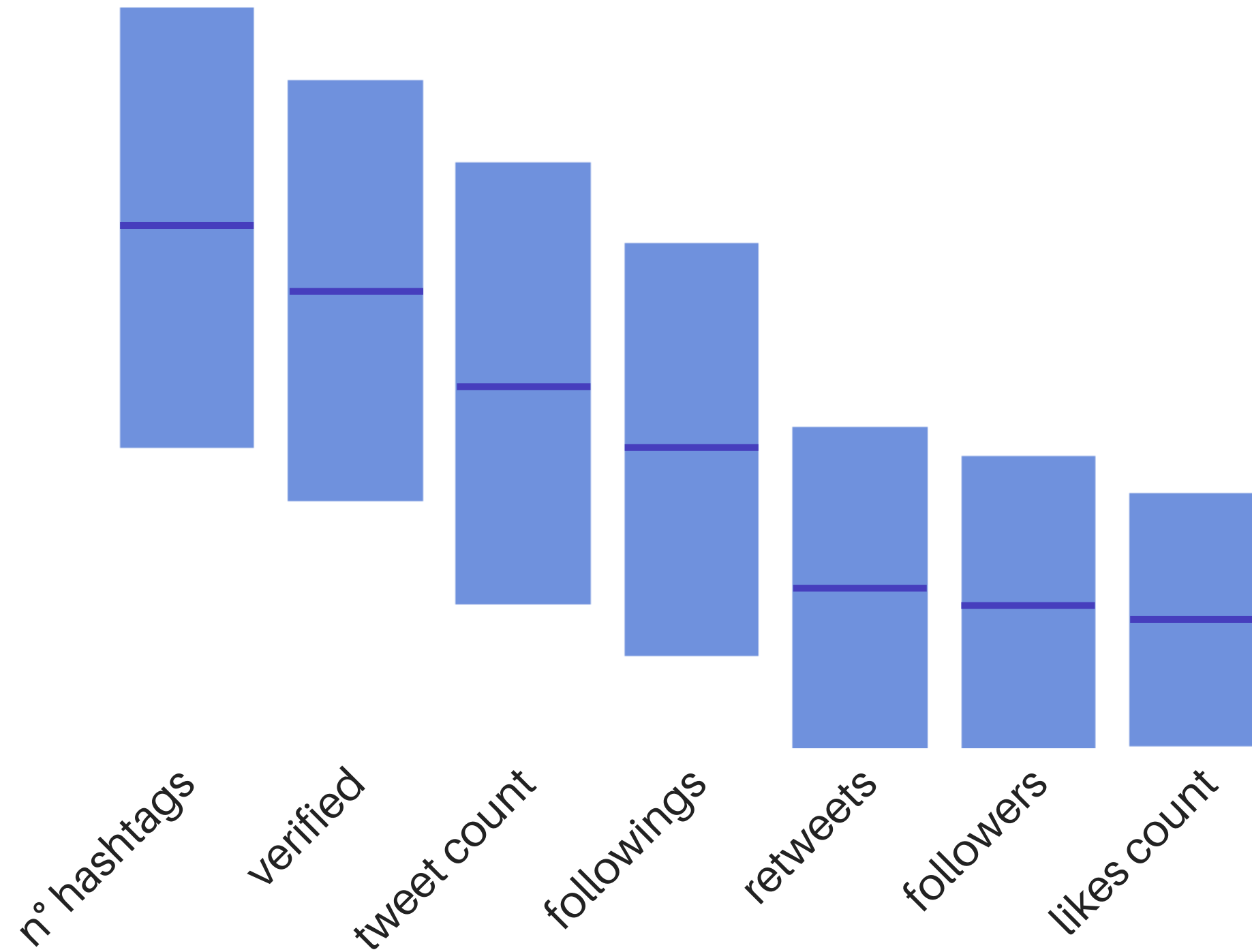
Accuracy per event

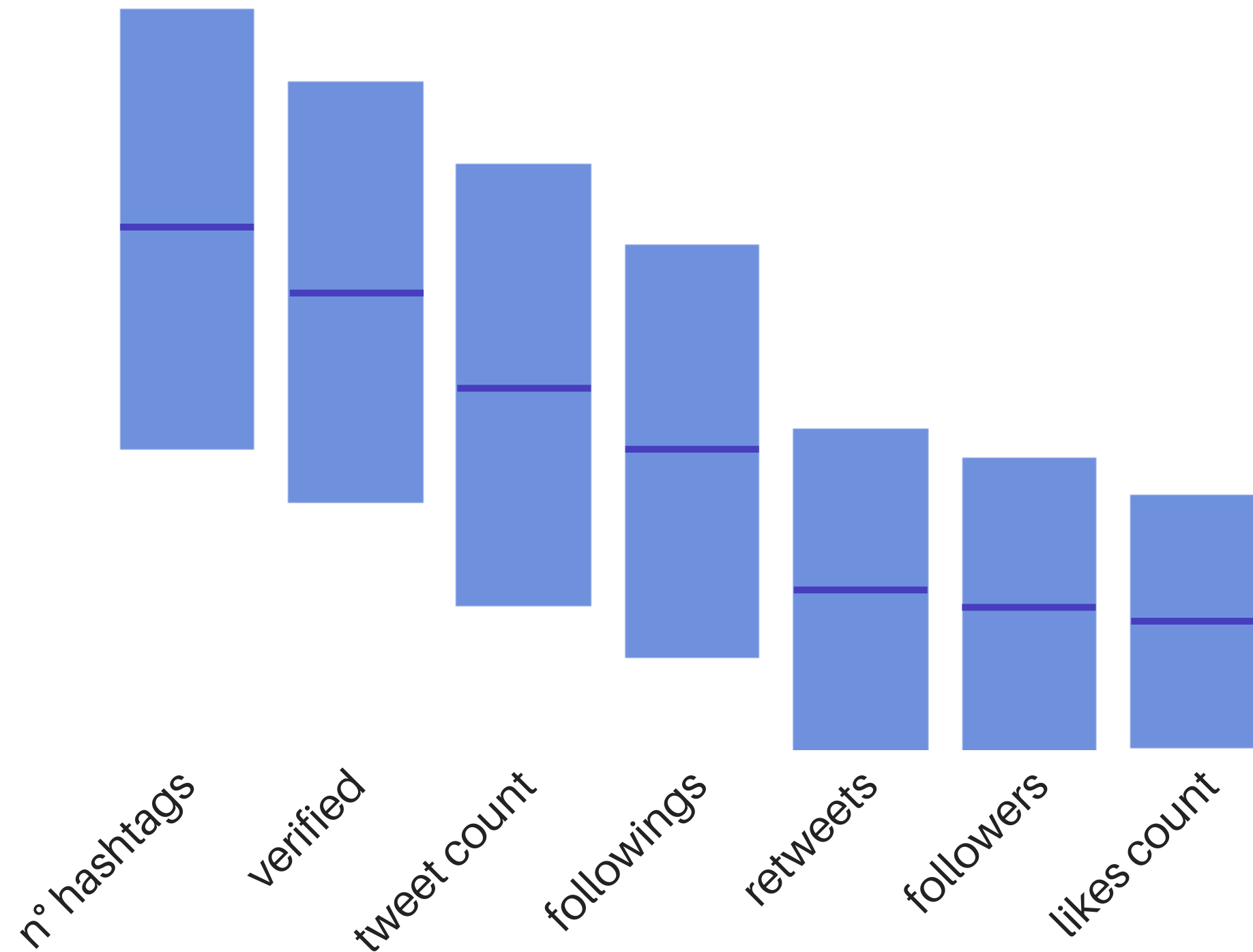Accuracy using Leave One Event Out as Cross Validation.

Feature Ranking

Feature Ranking

n° hashtags · verified · tweet count · followings · retweets · followers · likes count

Feature Ranking

Ranking can easily change between events

n° hashtags   verified   tweet count   followings   retweets   followers   likes count

DEPLOY OUR MODEL ON THE UKRANIAN WAR TWEETS

Keyword = "Ukraine"
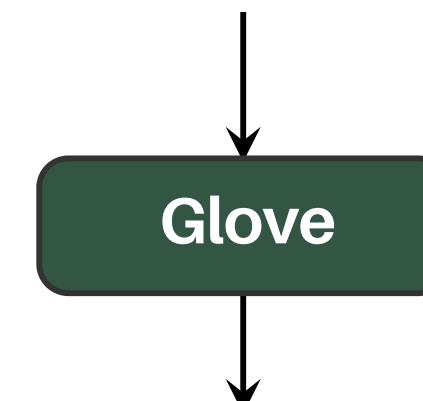
14 milions tweets gathered in March 2022
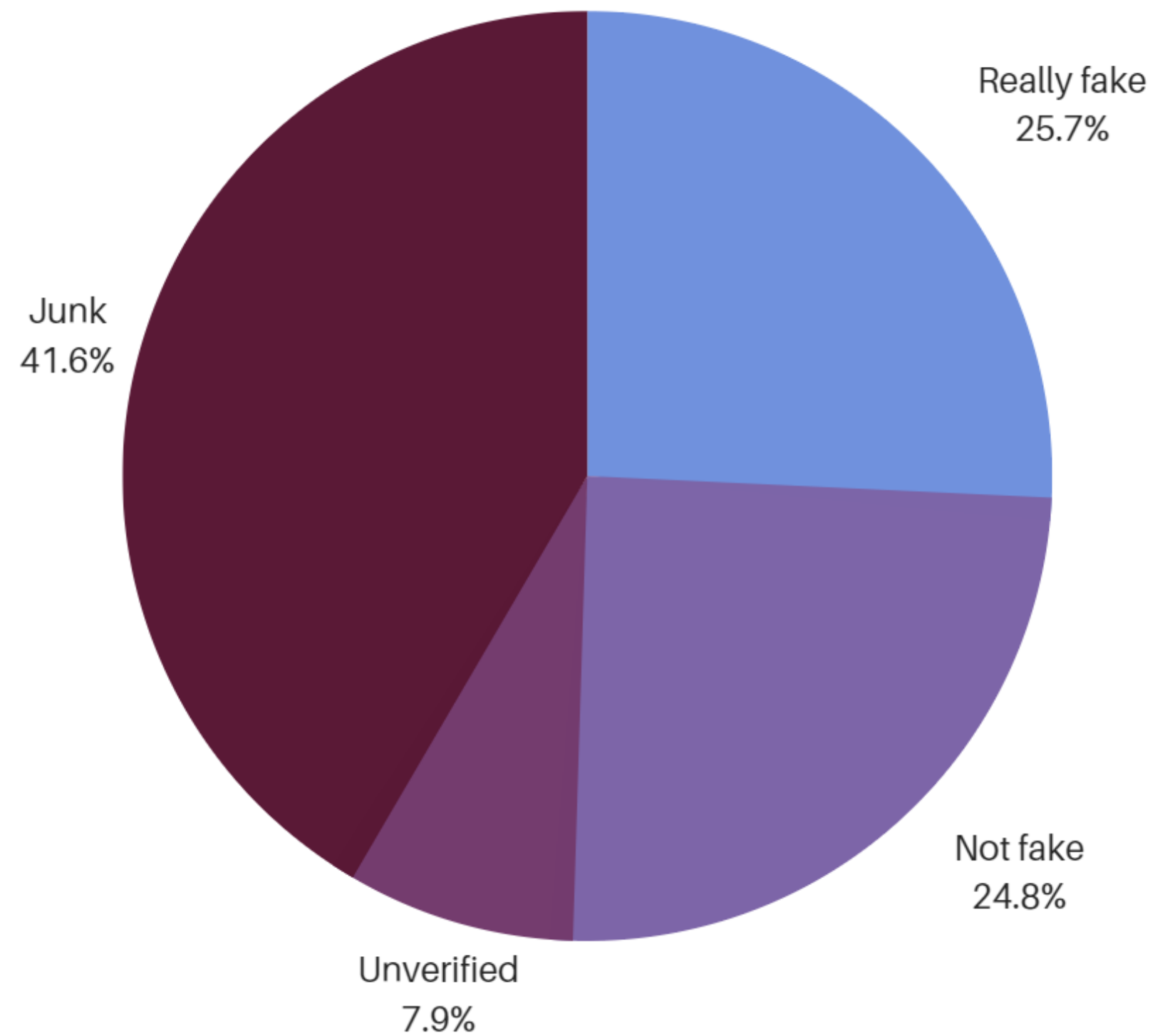
Datasets built

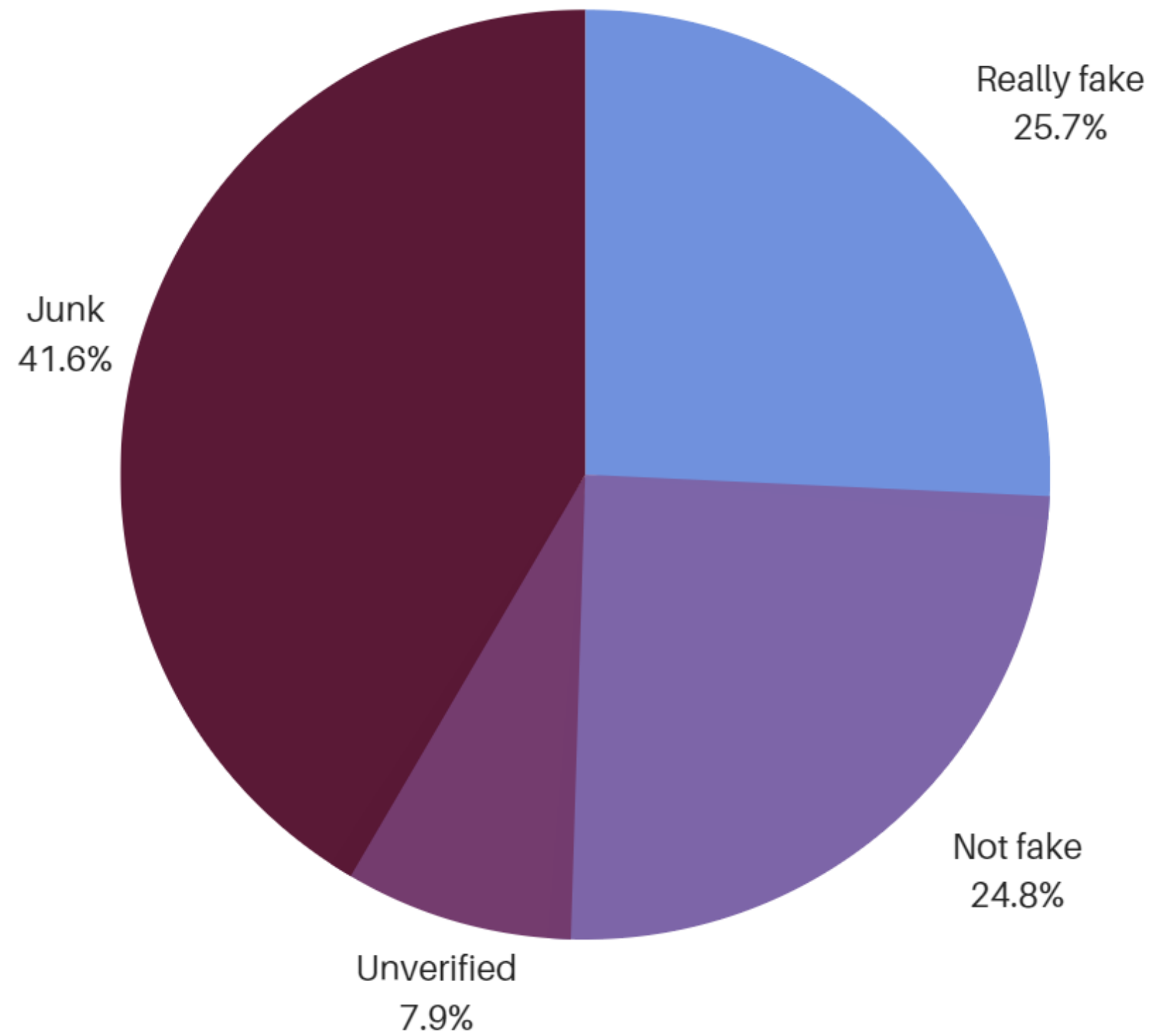RUS + ENG TWEETS

2223

↓

LASER

↓
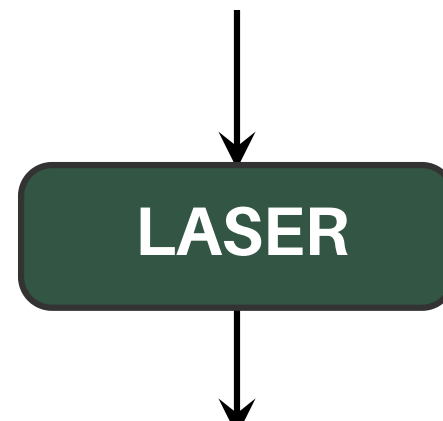
ENGLISH TWEETS

4000

↓

Glove

↓

Let's plug our model on Twitter

Russian and english tweets

Let's plug our model on Twitter

Checking the 65 most fakes

Really fake
25.7%

Not fake
24.8%

Unverified
7.9%

Junk
41.6%

15

Lots of junk

Let's try filter a bit

Junk
41.6%

Really fake
25.7%

Not fake
24.8%

Unverified
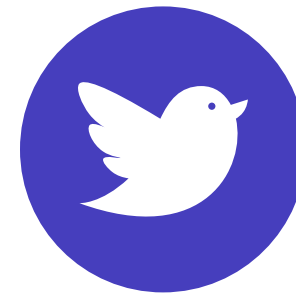7.9%

Really fake
9

Not fake
54
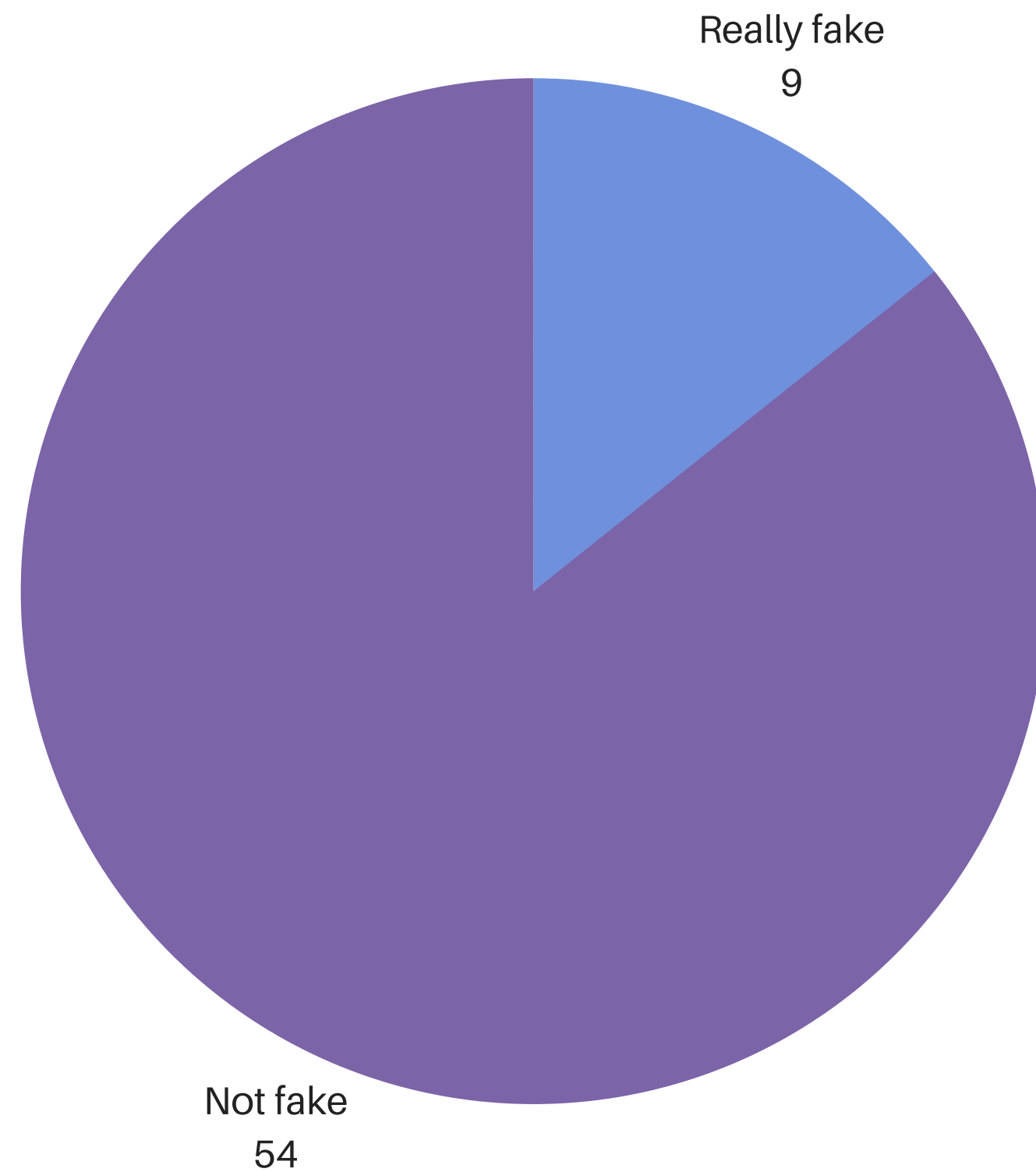
Filtered
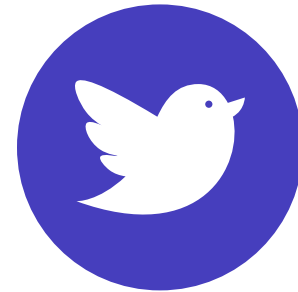# english tweets

Checking the 63 fakes

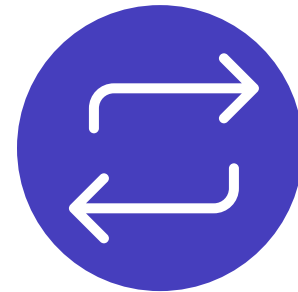16

Conclusions

Conclusions
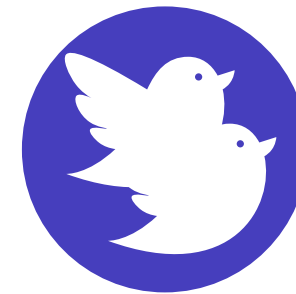
The single tweet it's not enough

Conclusions

The single tweet it's not enough

Big changes with different events

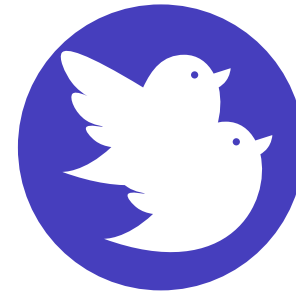# Can we reach that 82% ?

Can we reach that 82% ?

Look at the tweet replies

Can we reach that 82% ?

Look at the tweet replies

Use a dataset of claims

Fake News Detector

Eithar Abdalslam

Manuel Cecere

Antoine Luu

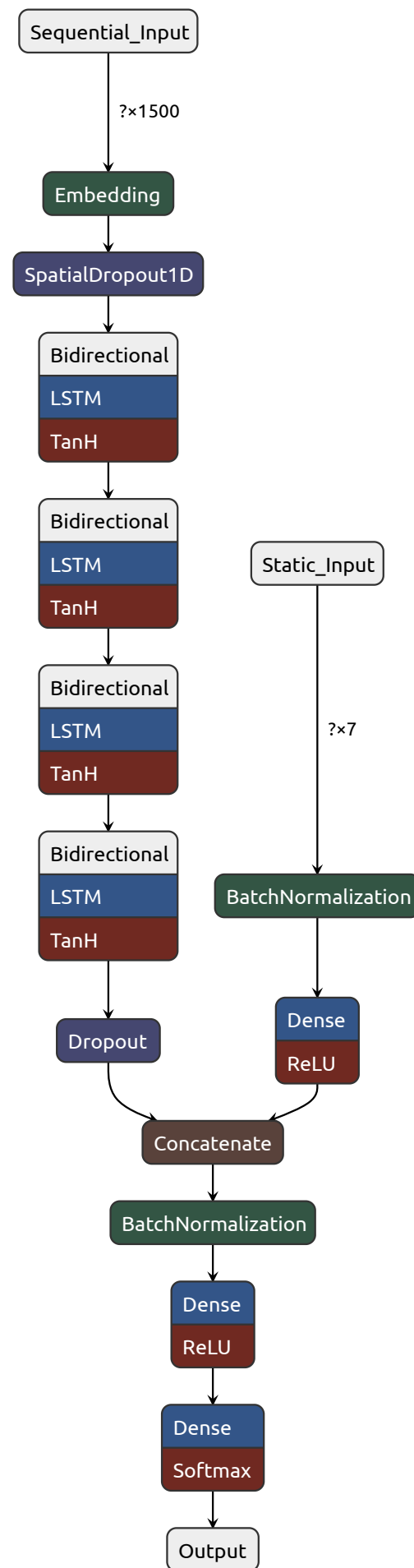Anastasia Cotov

19

Thanks for your time

20

Extra material

42

```
Sequential_Input

        │ ?×1500
        ▼
     Embedding

  SpatialDropout1D

    Bidirectional
        LSTM
        TanH

    Bidirectional
        LSTM
        TanH

    Bidirectional              Static_Input
        LSTM
        TanH                        │ ?×7
                                    ▼
    Bidirectional
        LSTM                 BatchNormalization
        TanH

      Dropout                     Dense
                                   ReLU

              Concatenate

           BatchNormalization

                 Dense
                  ReLU

                 Dense
                 Softmax

                Output
```
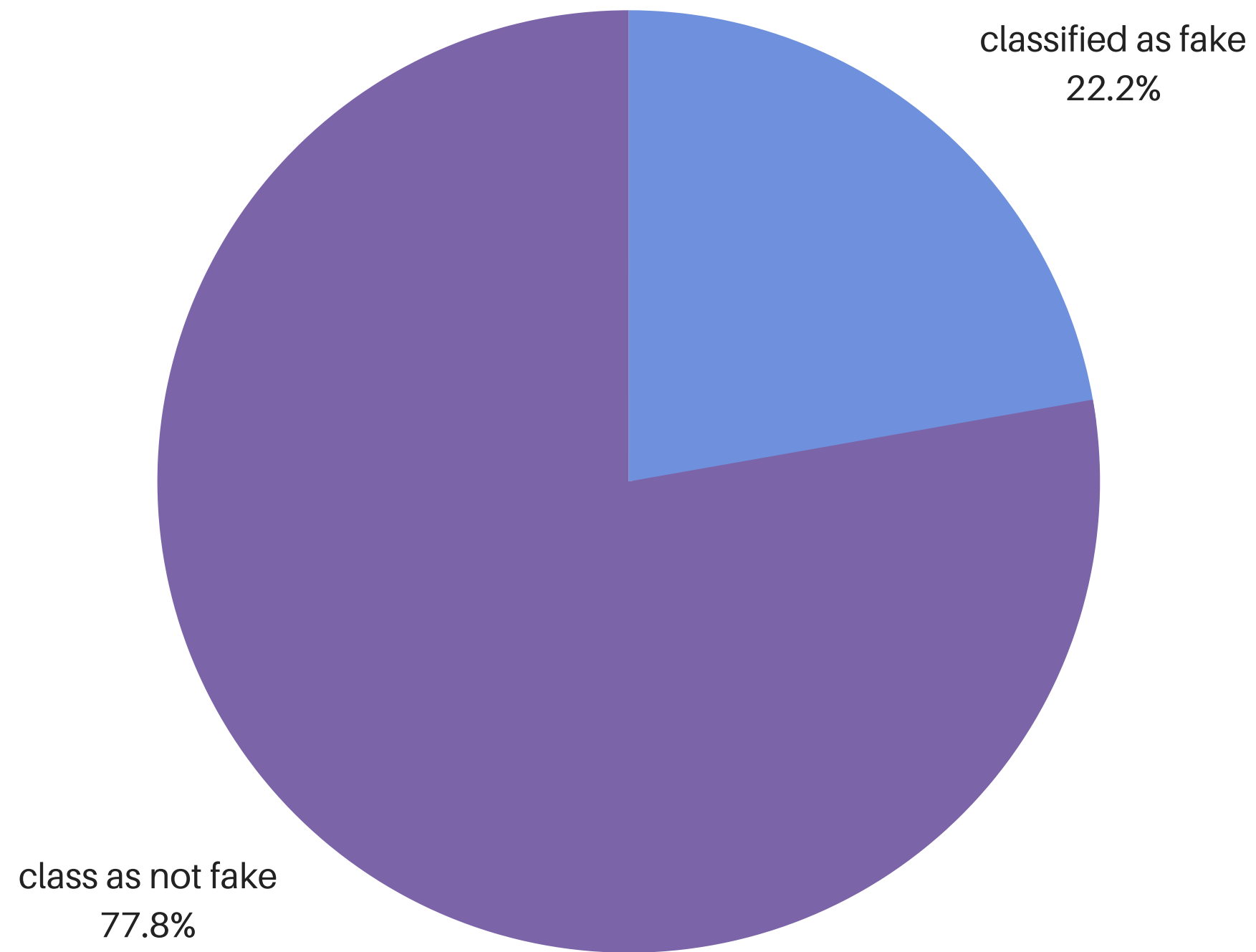
Final Glove

Final LASER

42

classified as fake
22.2%

class as not fake
77.8%
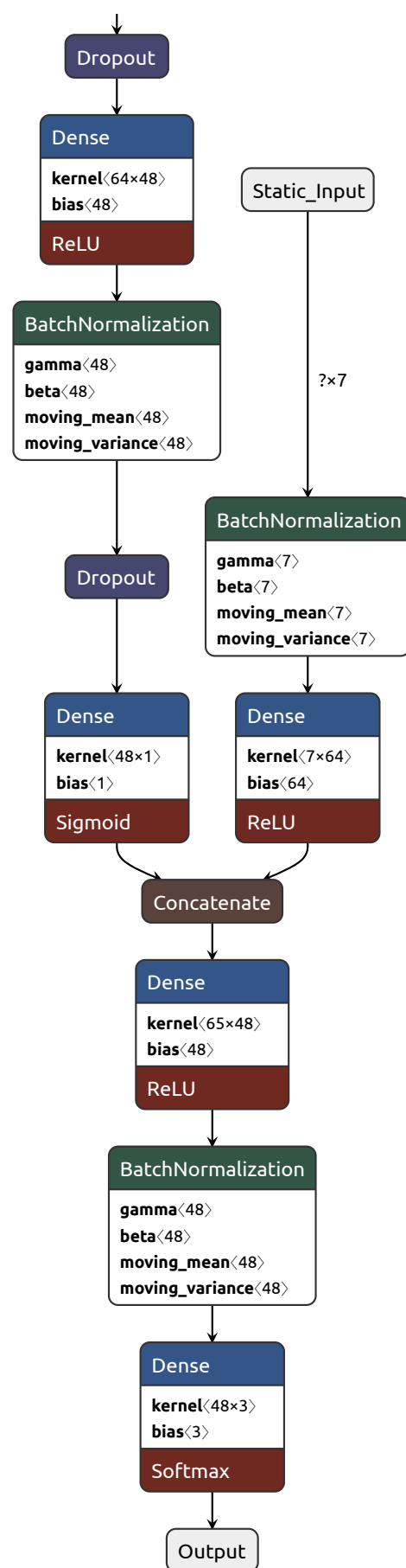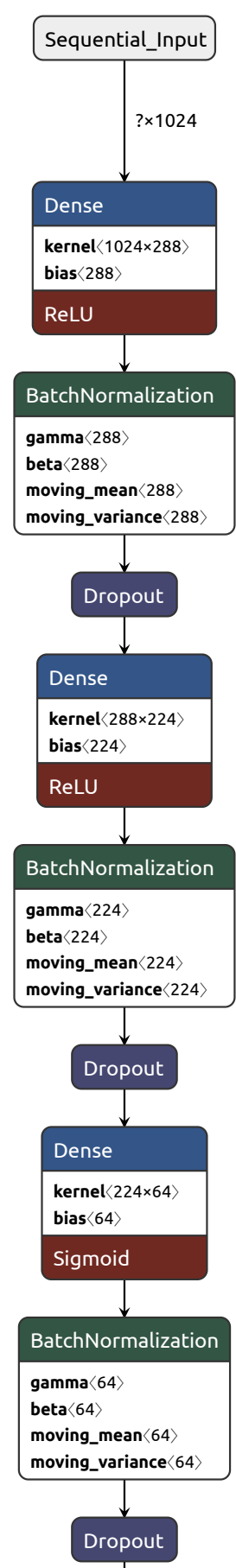
Obvious fake twees

A bit sad story

"Not fake is true or not fake is detected by the model as fake but they are not really fake?"

Eithar