Hate Speech Detection is Not as Easy as You May Think:

A Closer Look at Model Validation

Aymé Arango, Jorge Pérez and Bárbara Poblete

UNIVERSIDAD DE CHILE

Instituto Milenio Fundamentos de los datos

ALMOST PERFECT STATE-OF-THE-ART RESULTS

VS

UNDETECTED HATE SPEECH IN SOCIAL MEDIA

Bloomberg

Twitter Apologizes for Ignoring Bomb Suspect's Apparent Threat in Tweet

October 27, 2018, 12:04 AM GMT-3 *Updated on October 27, 2018, 1:20 AM GMT-3*

NEWS | RACIAL JUSTICE

Civil Rights Groups Have Been Warning Facebook About Hate Speech for Years

EdSurge

Twitter Is Funding Research Into Online Civility. Here's How One Project Will Work.

By Jeffrey R. Young Aug 14, 2018

THE UNIVERSITY OF SYDNEY

University researchers to help Facebook counter hate speech

30 May 2019

UNDETECTED HATE SPEECH IN SOCIAL MEDIA

ALMOST PERFECT STATE-OF-THE-ART RESULTS

94% F1
[Agrawal and Awekar]
ECIR
2018

93% F1
[Badjatiya et al.]
WWW
2017

92% F1
[Zeerak Waseem]
NAACL
2016

Hate Speech Detection is Not as Easy as You May Think

We show that state of the art results are highly overestimated due to experimental issues in the models:

Including the testing set during training phase

Oversampling the data before splitting

User-biased datasets

State-of-the-art replication

User distribution

Generalization

State-of-the-art replication

User distribution

Generalization

ALMOST PERFECT STATE-OF-THE-ART RESULTS

94% F1
[Agrawal and Awekar]
ECIR
2018

93% F1
[Badjatiya et al.]
WWW
2017

92% F1
[Zeerak Waseem]
NAACL
2016

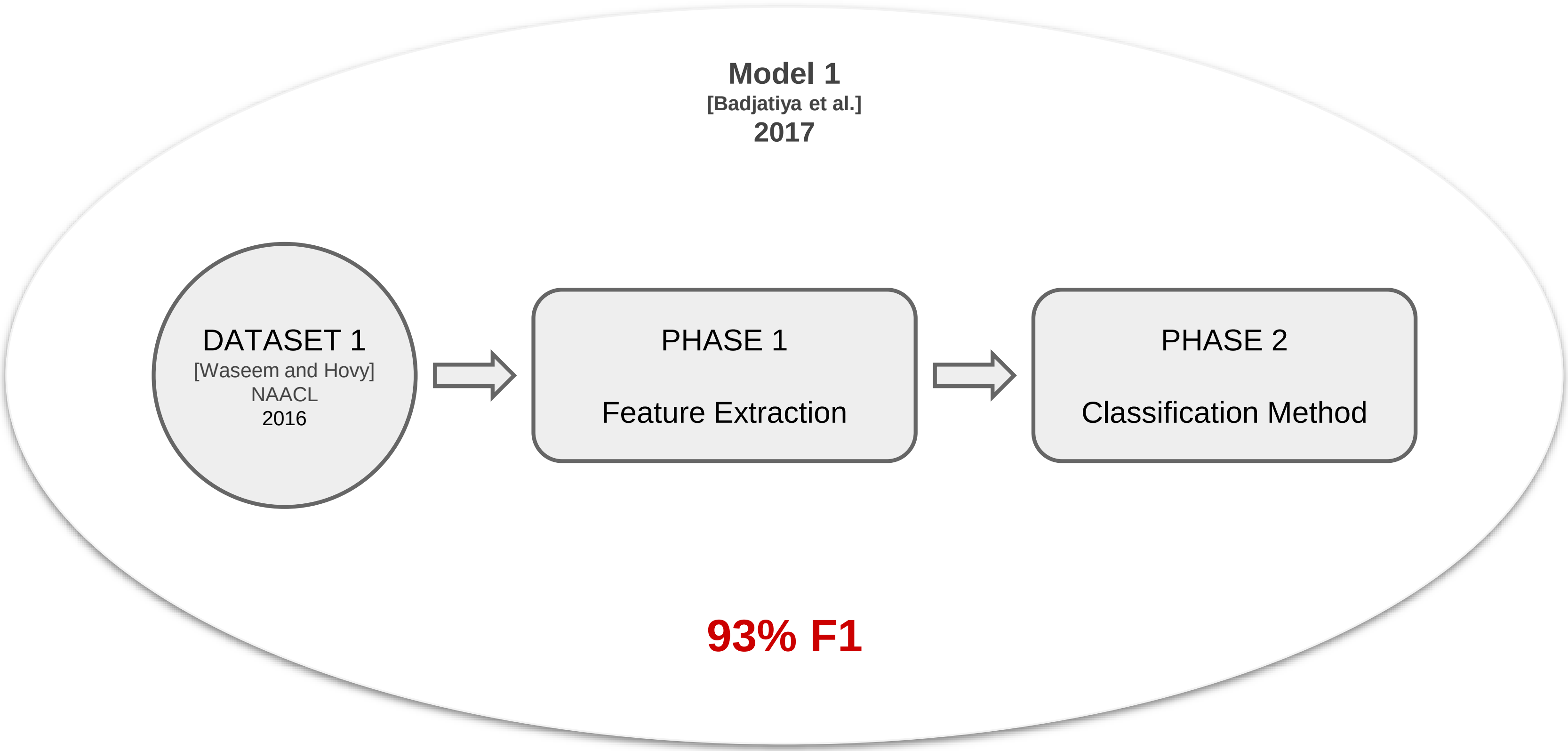
DATASET 1

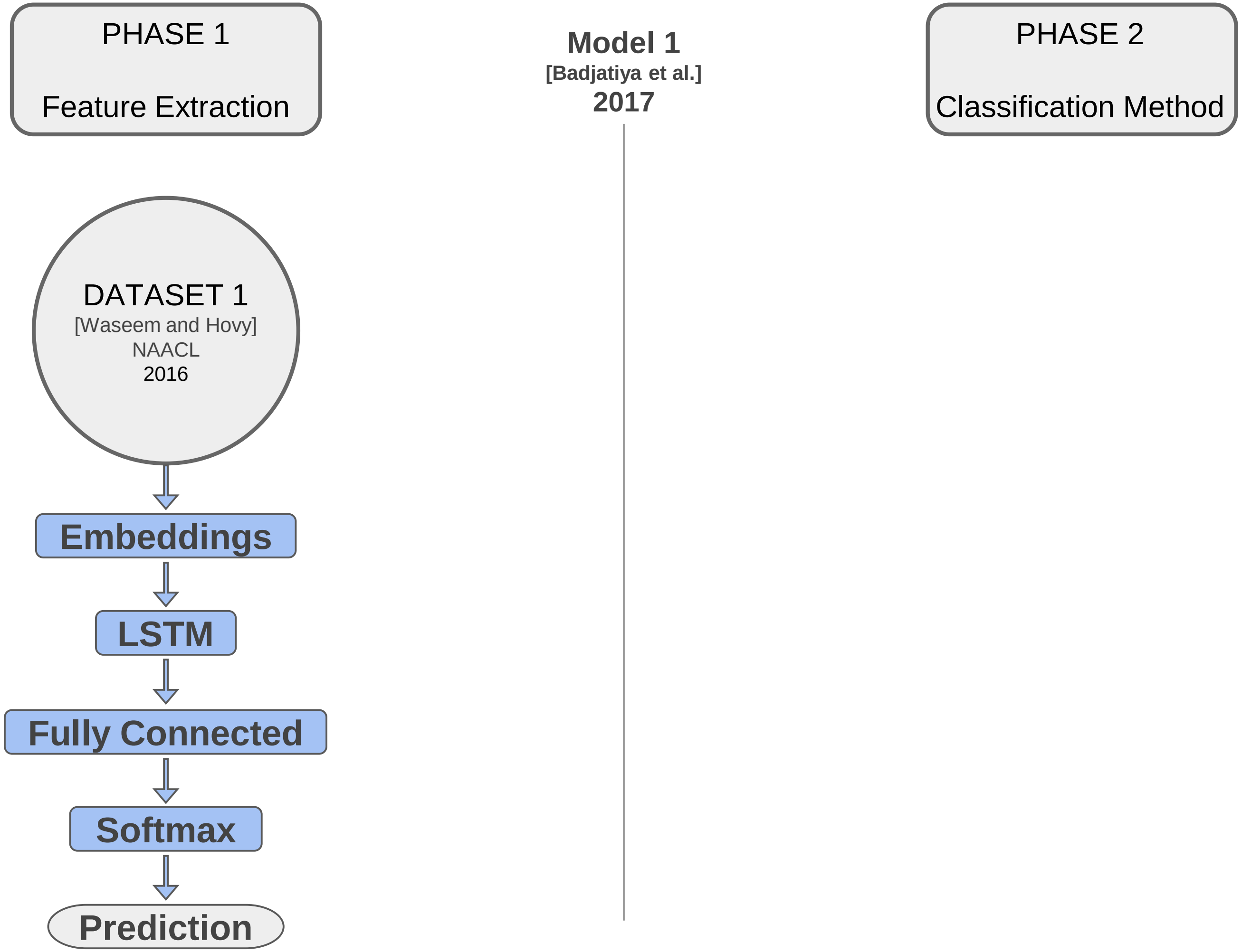
[Waseem and Hovy]

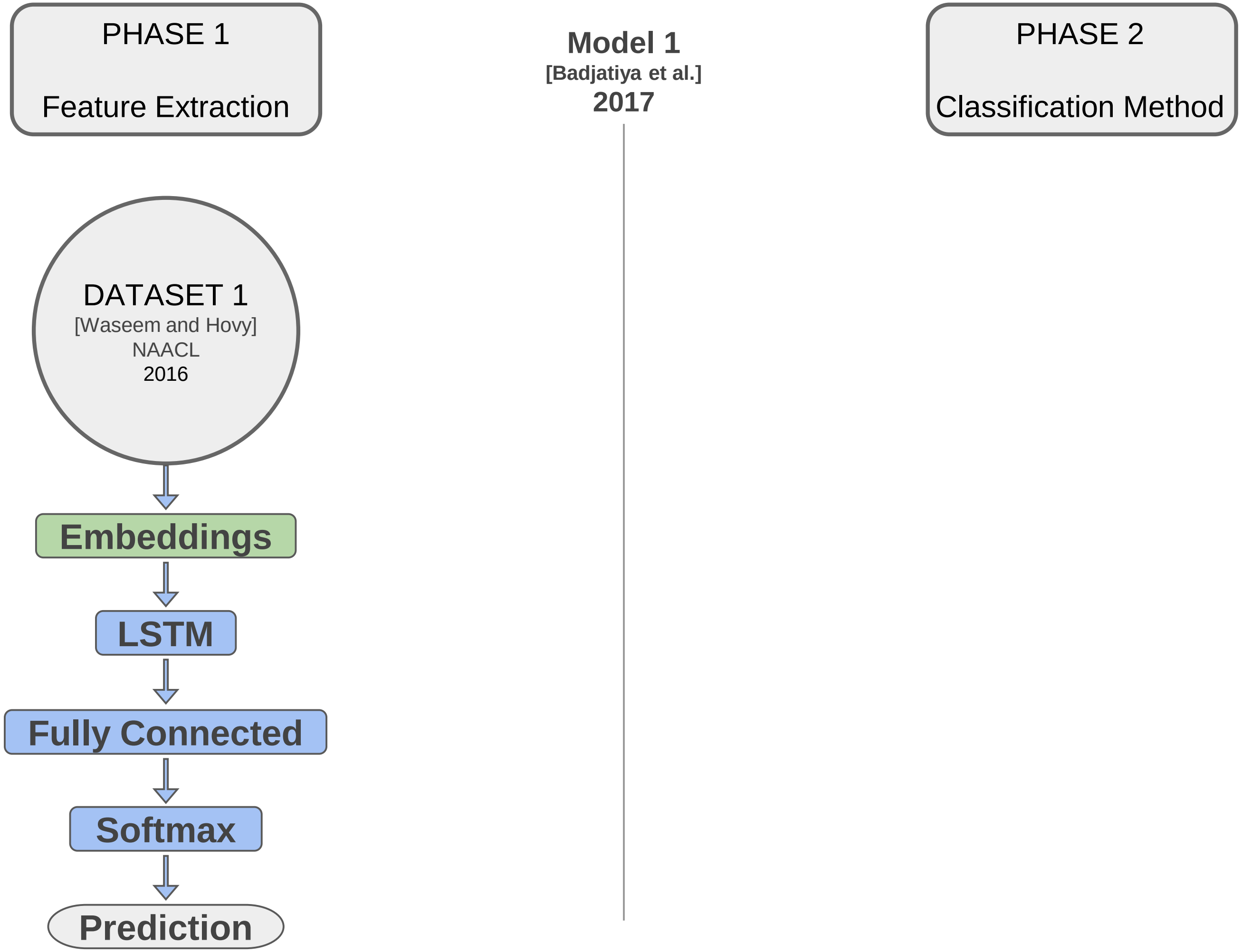
NAACL

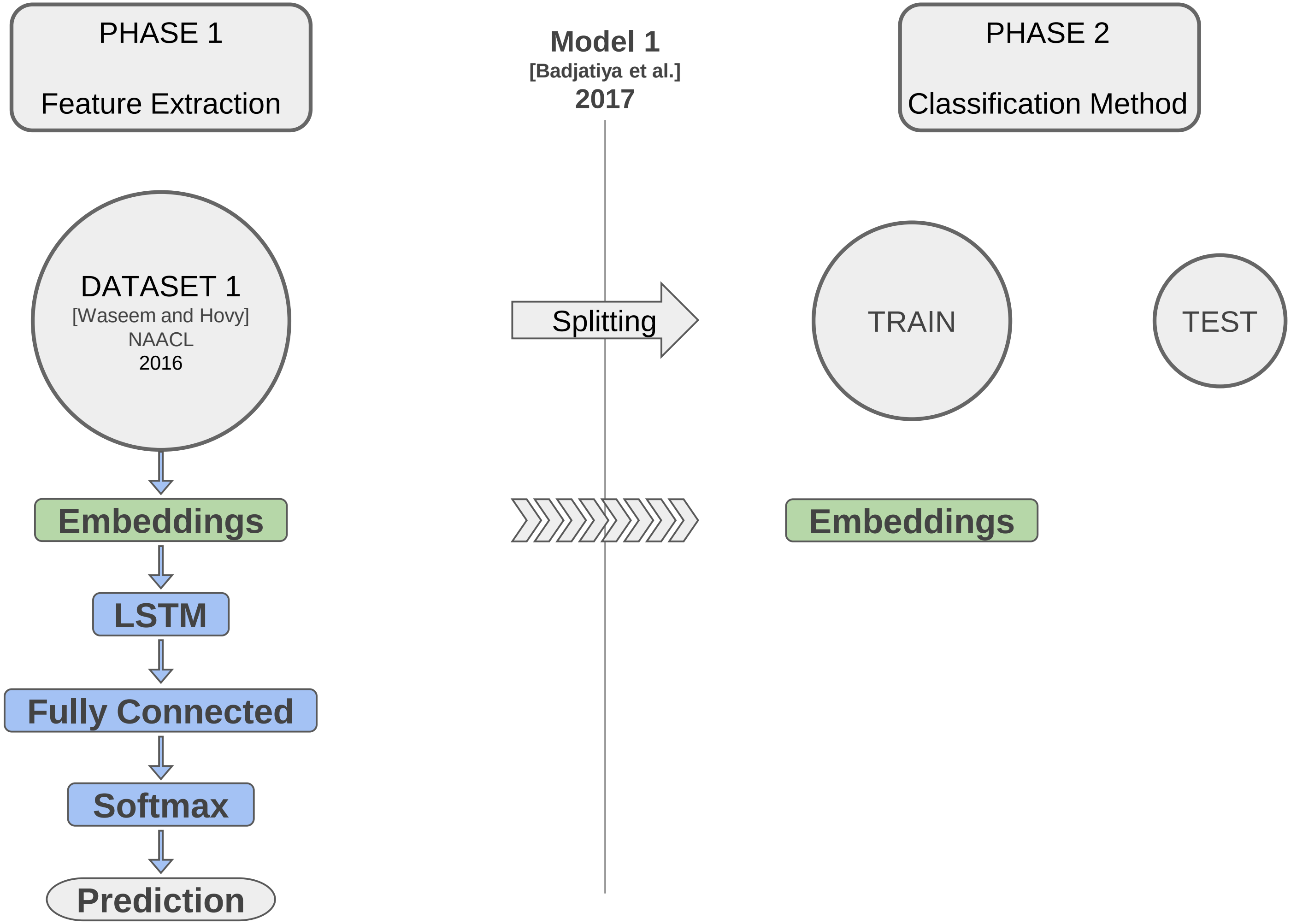
2016

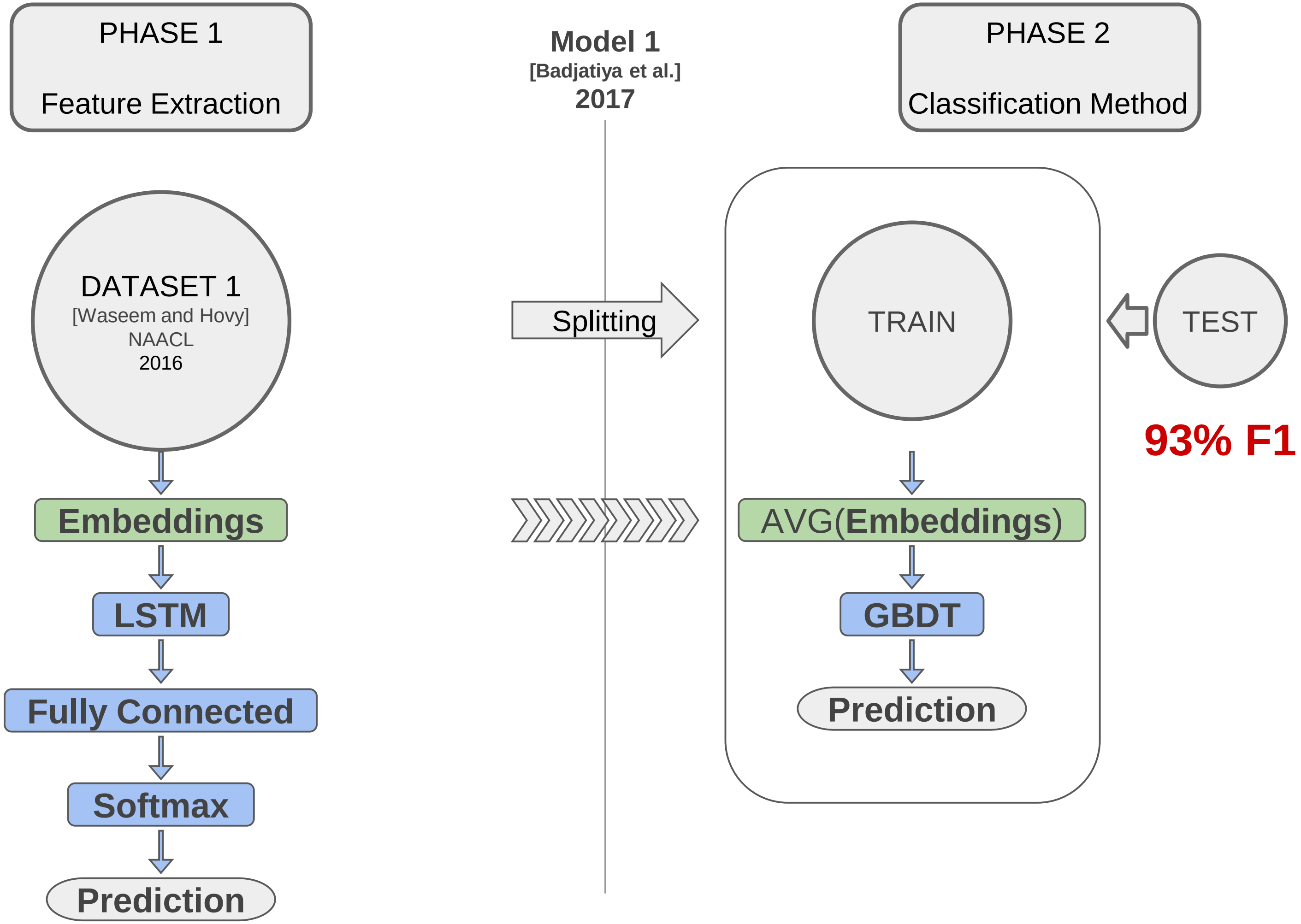
Tweet	Label
	Hate
	Non-Hate


Model 1
[Badjatiya et al.]
2017
DATASET 1
[Waseem and Hovy]
NAACL
2016
PHASE 1
Feature Extraction
PHASE 2
Classification Method
93% F1

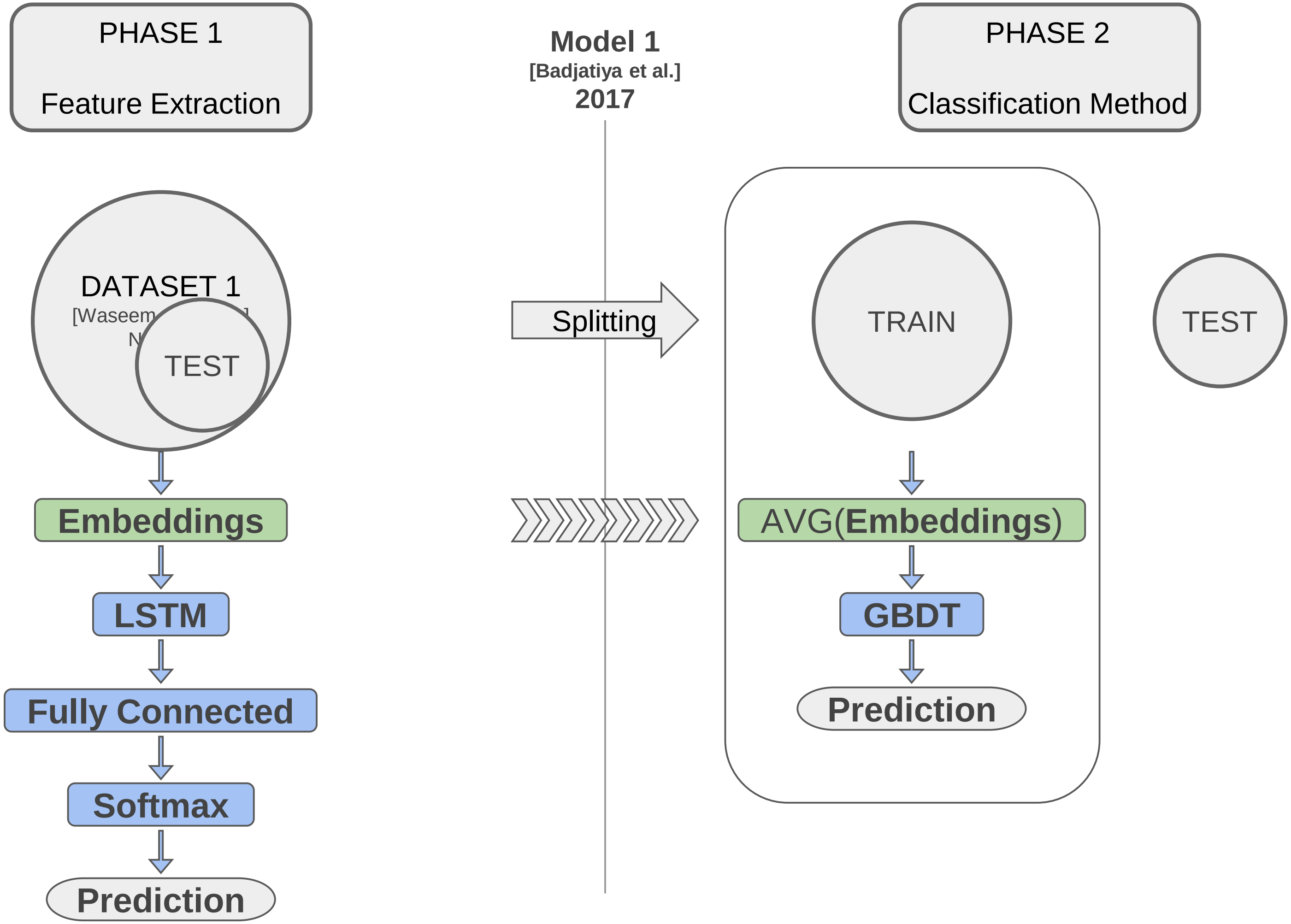

PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
DATASET 1
[Waseem and Hovy]
NAACL
2016
Embeddings
LSTM
Fully Connected
Softmax
Prediction


PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
DATASET 1
[Waseem and Hovy]
NAACL
2016
Embeddings
LSTM
Fully Connected
Softmax
Prediction


PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
DATASET 1
[Waseem and Hovy]
NAACL
2016
Splitting
TRAIN
TEST
Embeddings
Embeddings
LSTM
Fully Connected
Softmax
Prediction


PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
DATASET 1
[Waseem and Hovy]
NAACL
2016
Splitting
TRAIN
TEST
93% F1
Embeddings
AVG(Embeddings)
LSTM
GBDT
Fully Connected
Prediction
Softmax
Prediction

This looks great! But there is a problem.


PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
DATASET 1
TEST
Splitting
TRAIN
TEST
Embeddings
LSTM
Fully Connected
Softmax
Prediction
AVG(Embeddings)
GBDT
Prediction

Let’s create the model only with the training set.

PHASE 1

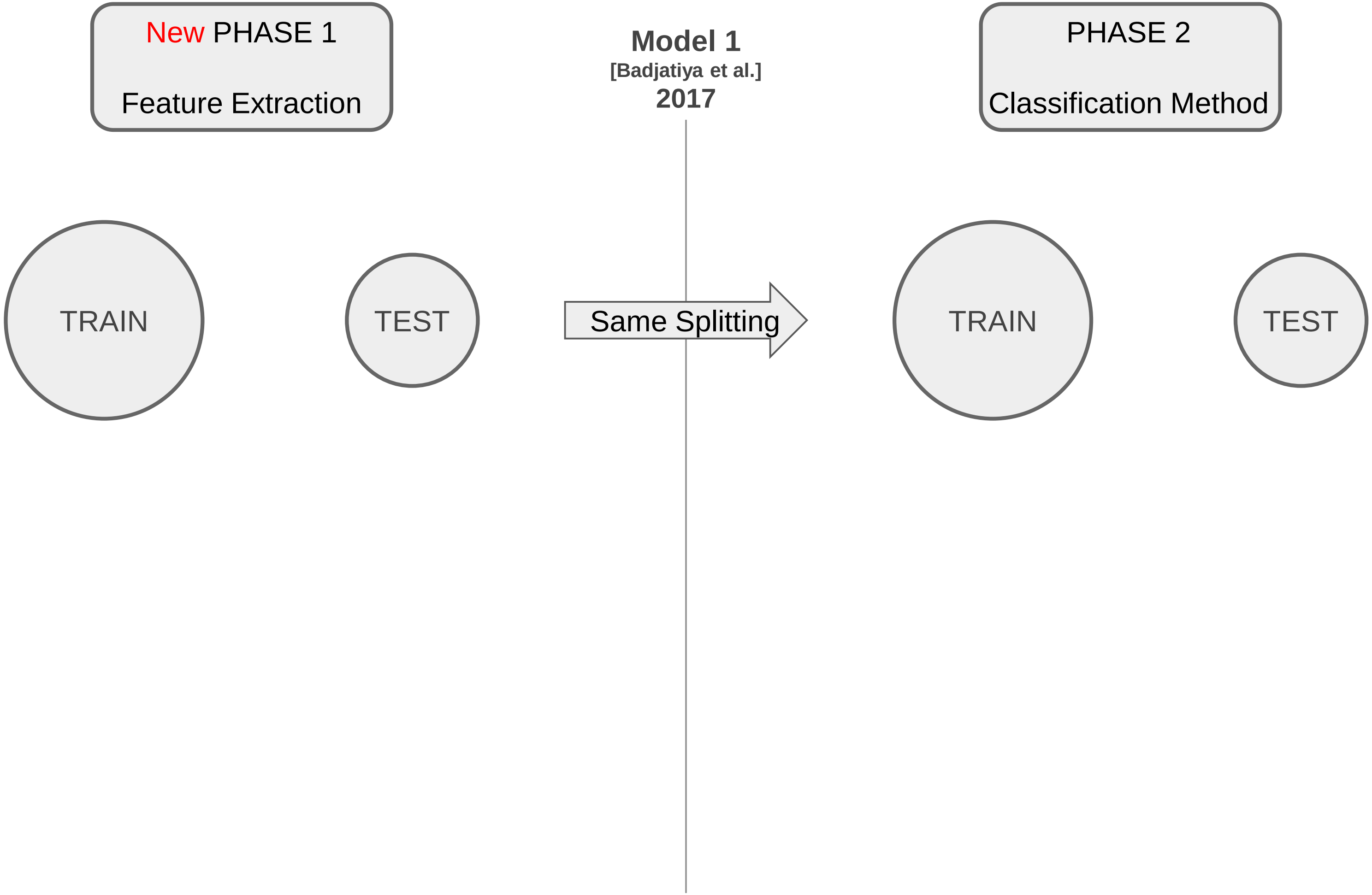
Feature Extraction

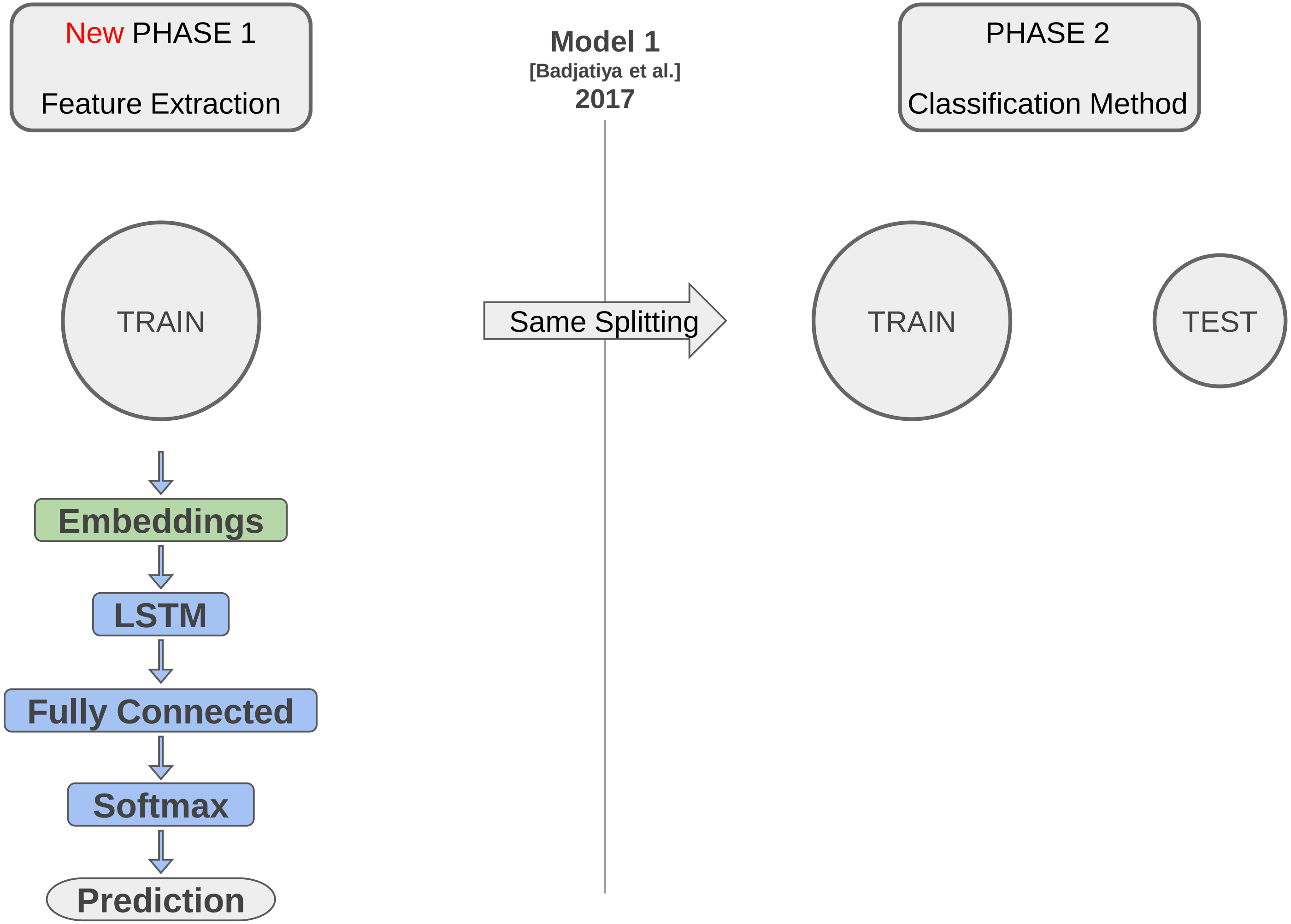
Model 1
[Badjatiya et al.]
2017

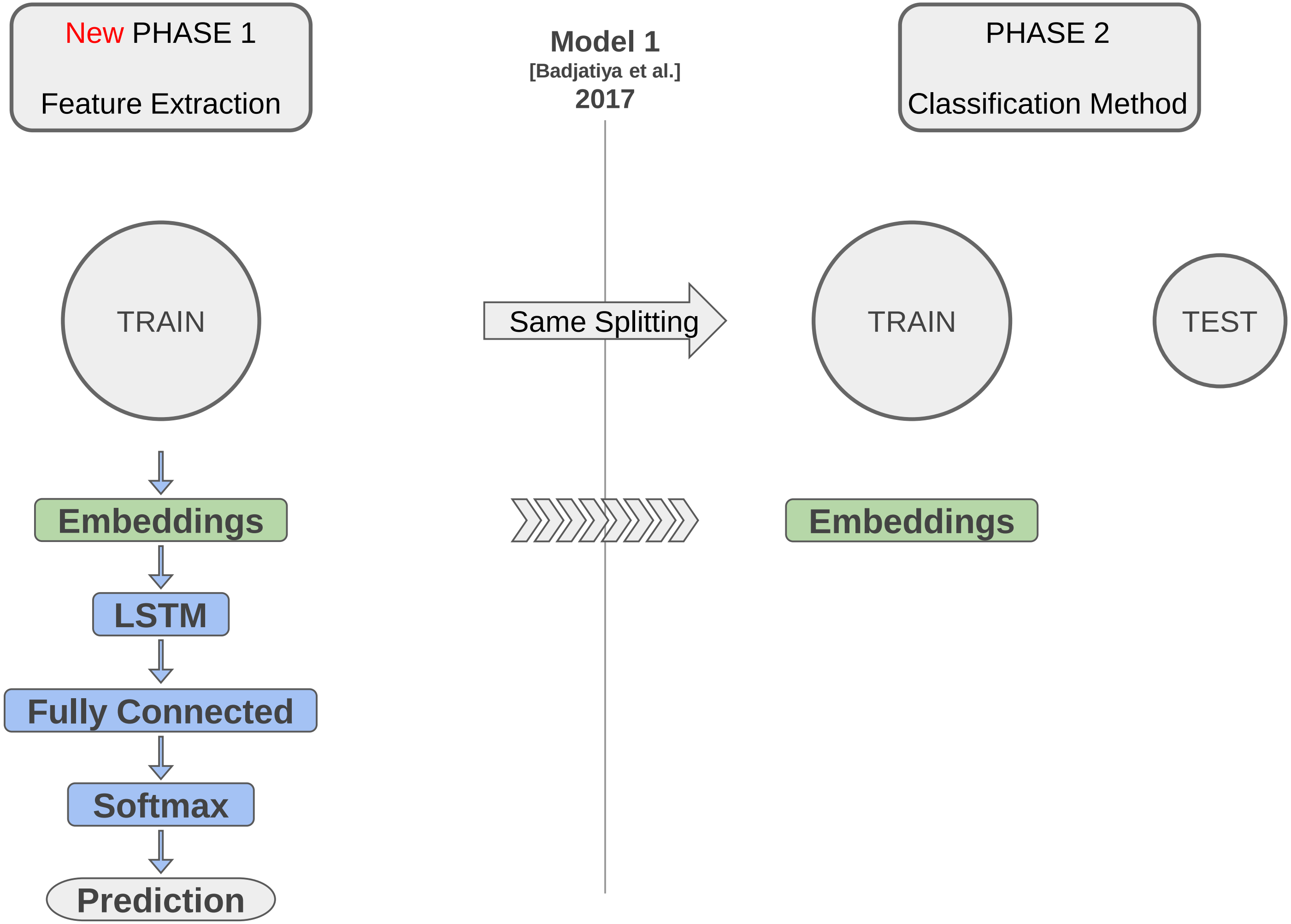
PHASE 2

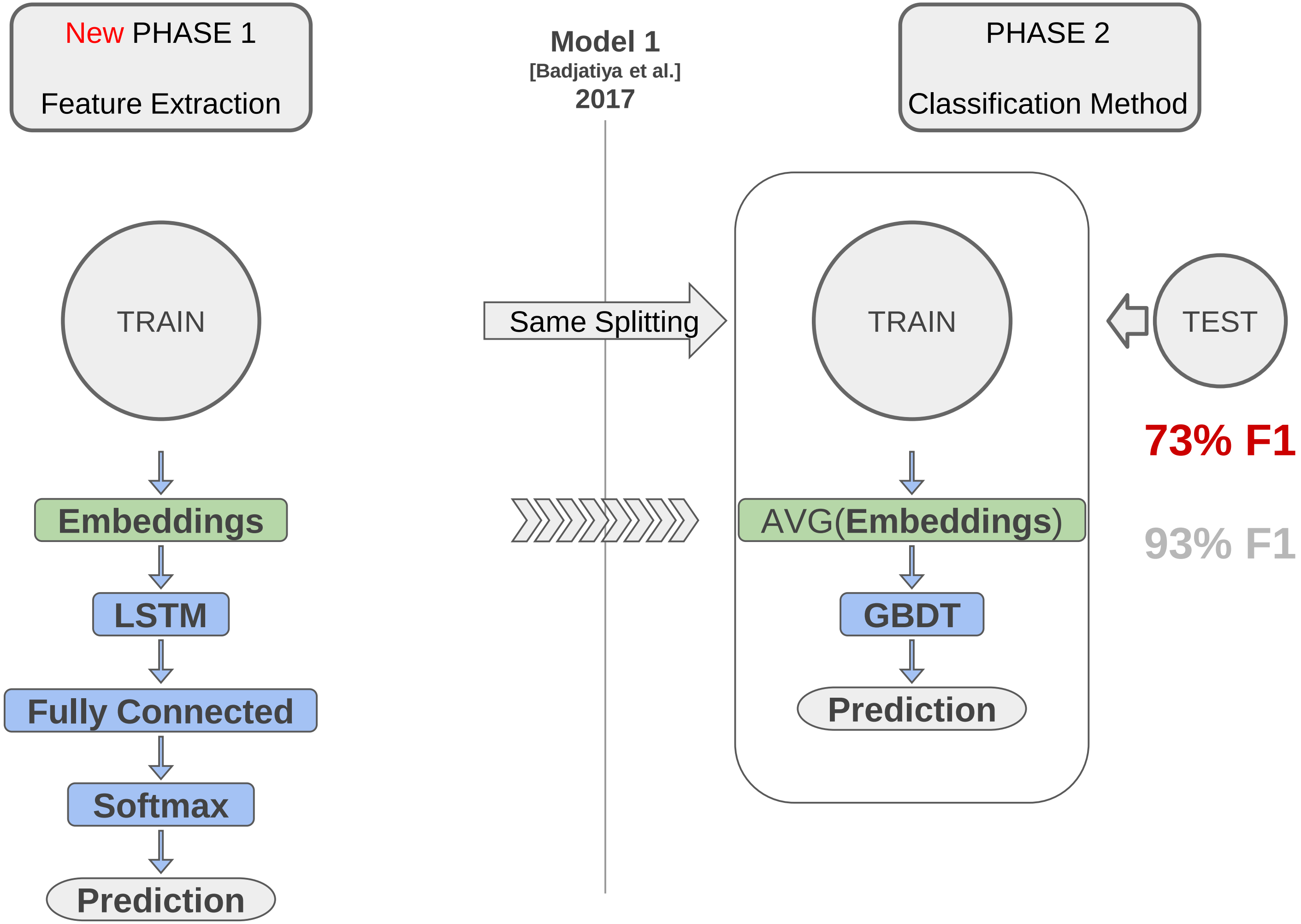
Classification Method

DATASET 1
[Waseem and Hovy]
NAACL
2016

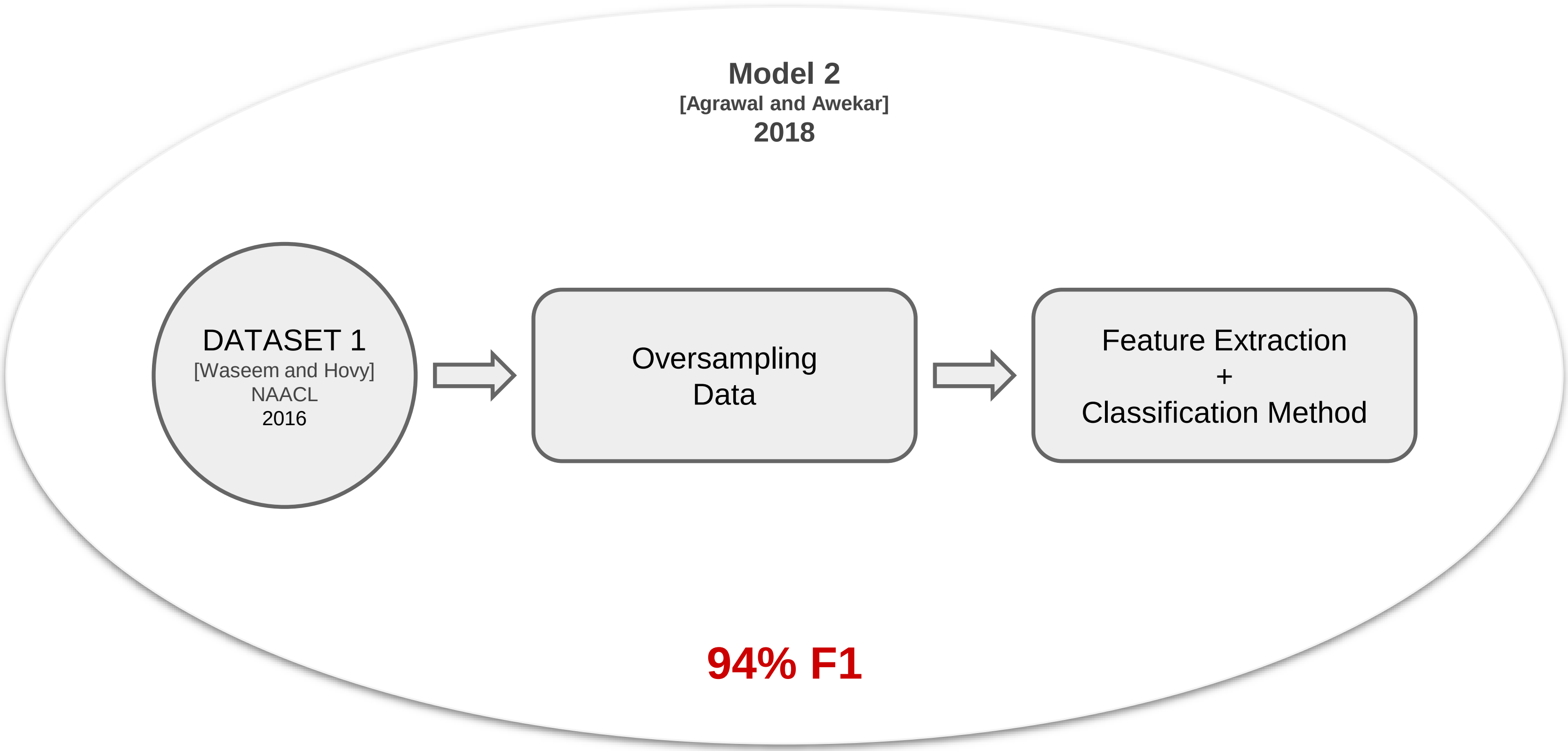

New PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
TRAIN
TEST
Same Splitting
TRAIN
TEST


New PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
TRAIN
Same Splitting
TRAIN
TEST
Embeddings
LSTM
Fully Connected
Softmax
Prediction


New PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
TRAIN
Same Splitting
TRAIN
TEST
Embeddings
Embeddings
LSTM
Fully Connected
Softmax
Prediction


New PHASE 1
Feature Extraction
Model 1
[Badjatiya et al.]
2017
PHASE 2
Classification Method
TRAIN
Same Splitting
TRAIN
TEST
73% F1
Embeddings
AVG(Embeddings)
93% F1
LSTM
GBDT
Fully Connected
Prediction
Softmax
Prediction

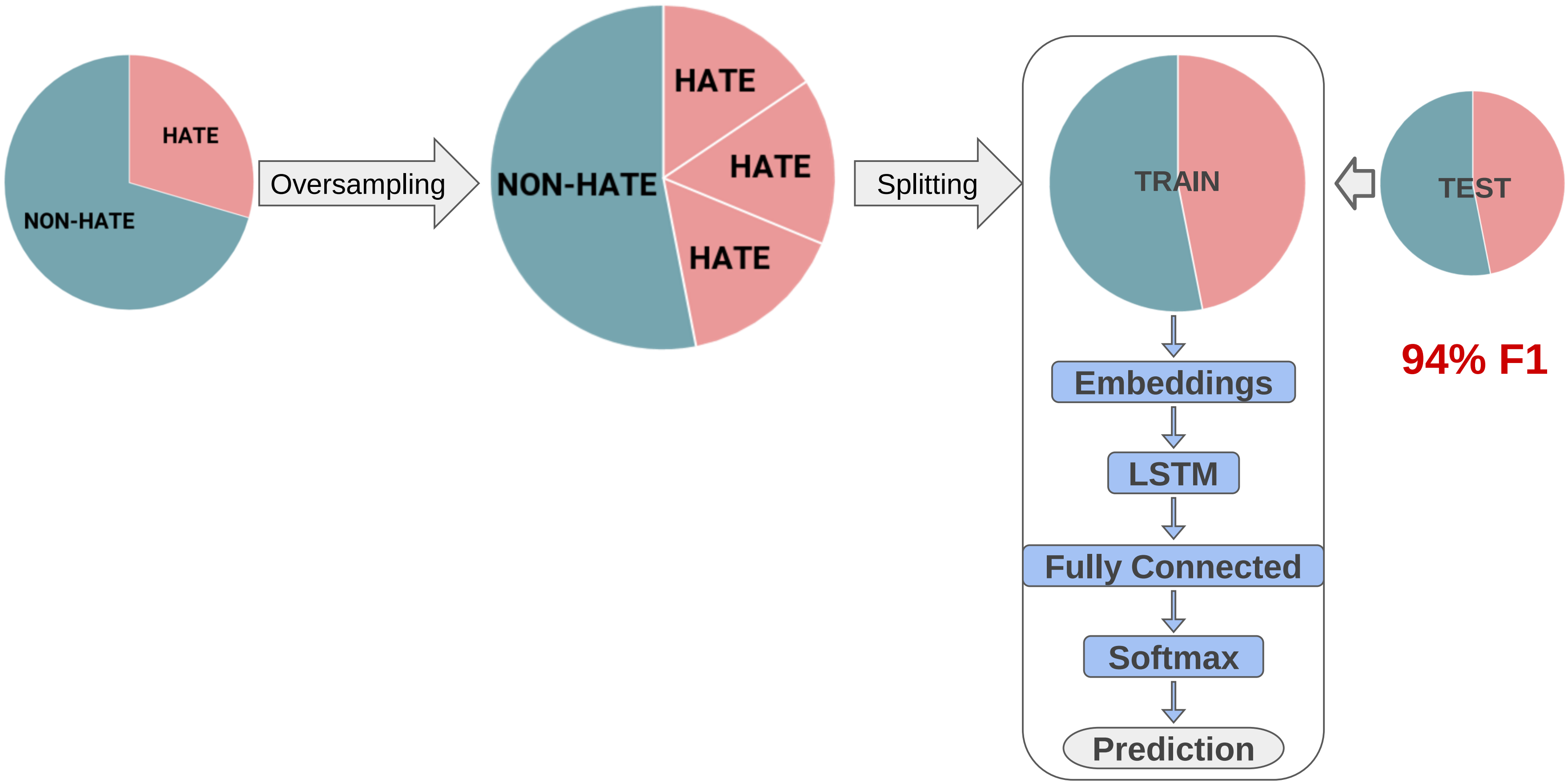
The result is overestimated due to
the inclusion of the testing set during the training phase.


Model 2
[Agrawal and Awekar]
2018
DATASET 1
[Waseem and Hovy]
NAACL
2016
Oversampling
Data
Feature Extraction
+
Classification Method
94% F1

Model 2
[Agrawal and Awekar]
2018

DATASET 1
[Waseem and Hovy]
NAACL
2016

Model 2
[Agrawal and Awekar]
2018



This also looks great! But there is another problem.

Model 2

[Agrawal and Awekar]

2018

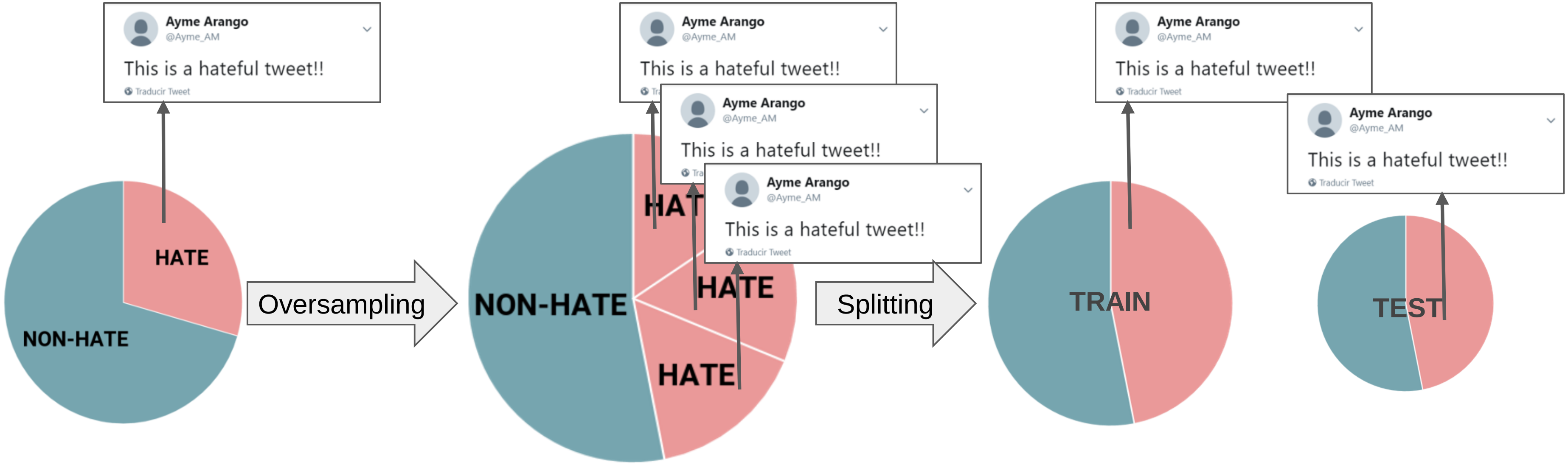
DATASET 1

[Waseem and Hovy]

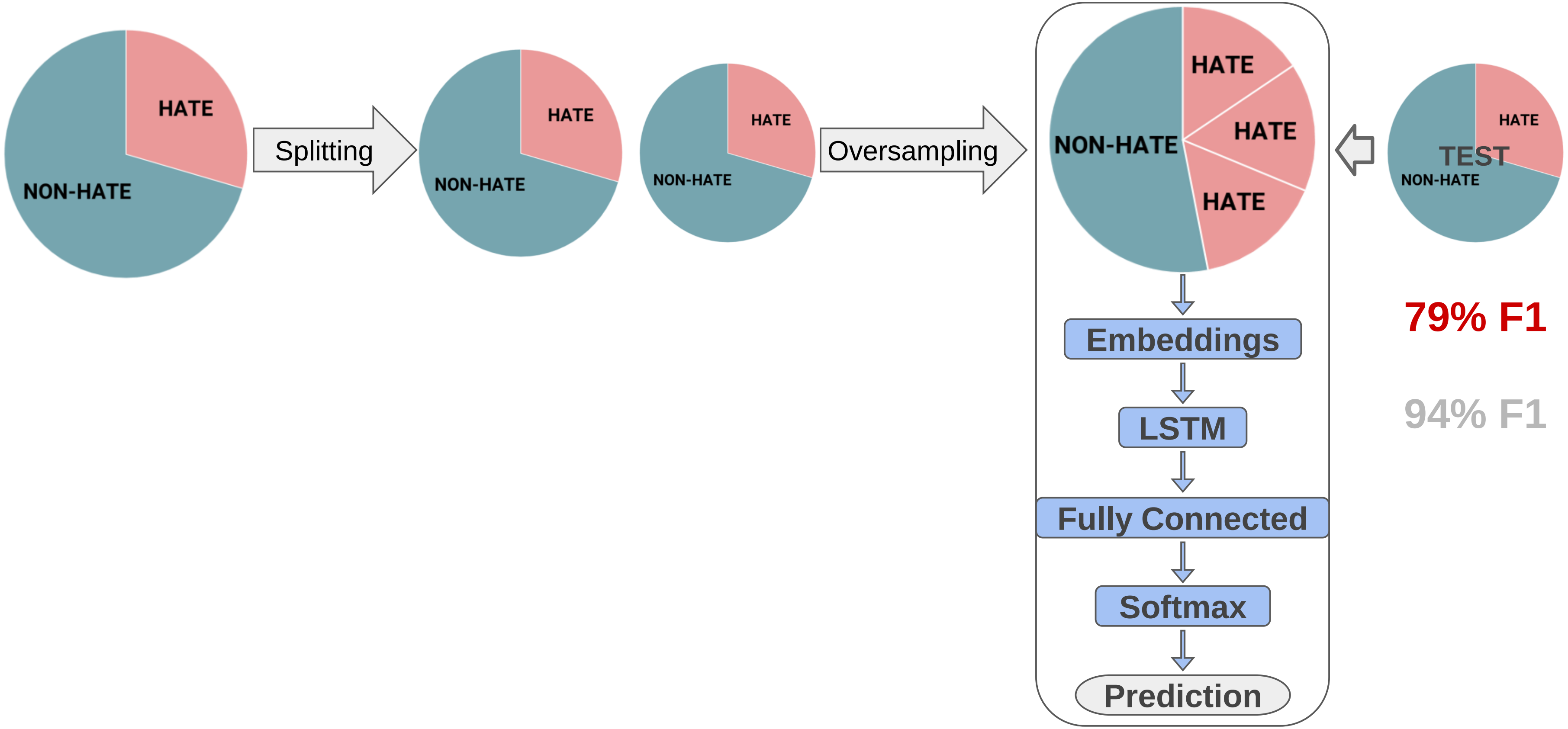
NAACL

2016

Model 2
[Agrawal and Awekar]
2018



Model 2
[Agrawal and Awekar]
2018



79% F1

94% F1

The result is overestimated due to the fact that the oversampling phase occurs before splitting the data.

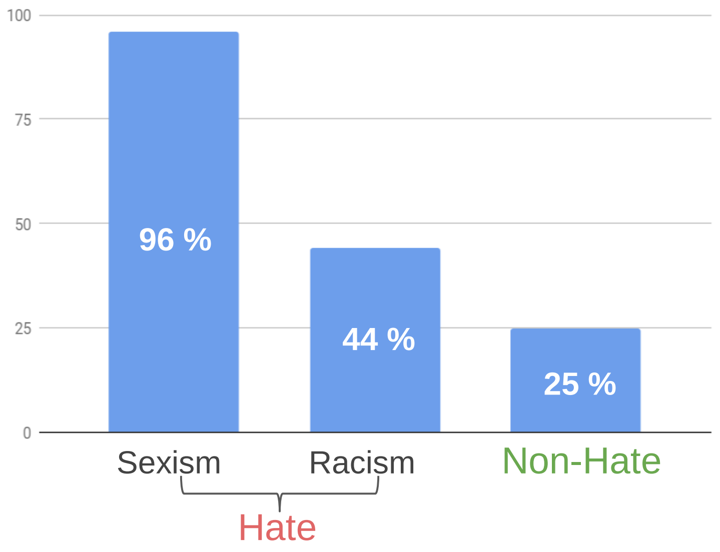
However, there is another issue to take into account.

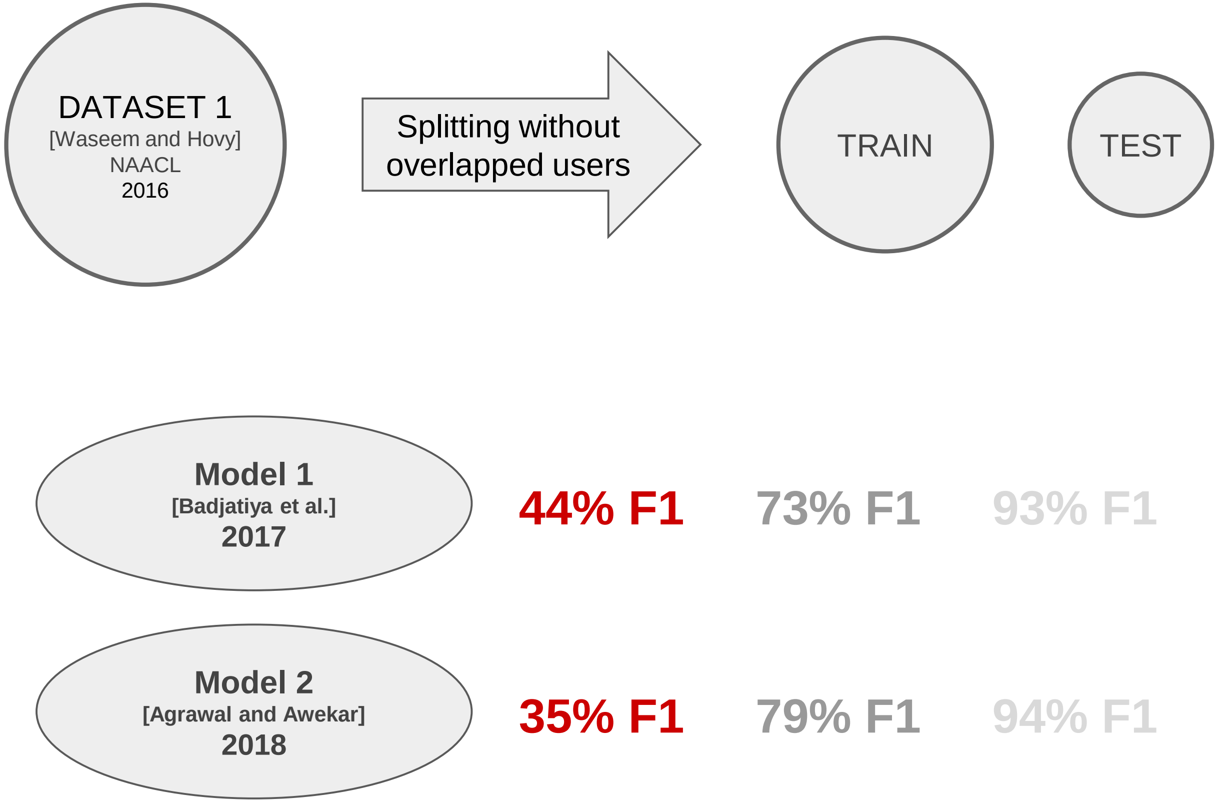
State-of-the-art replication

User distribution

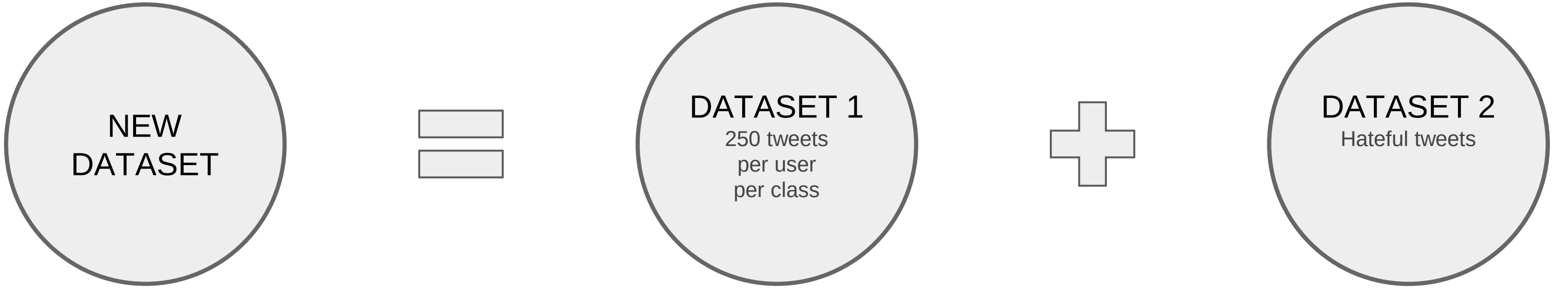
Generalization

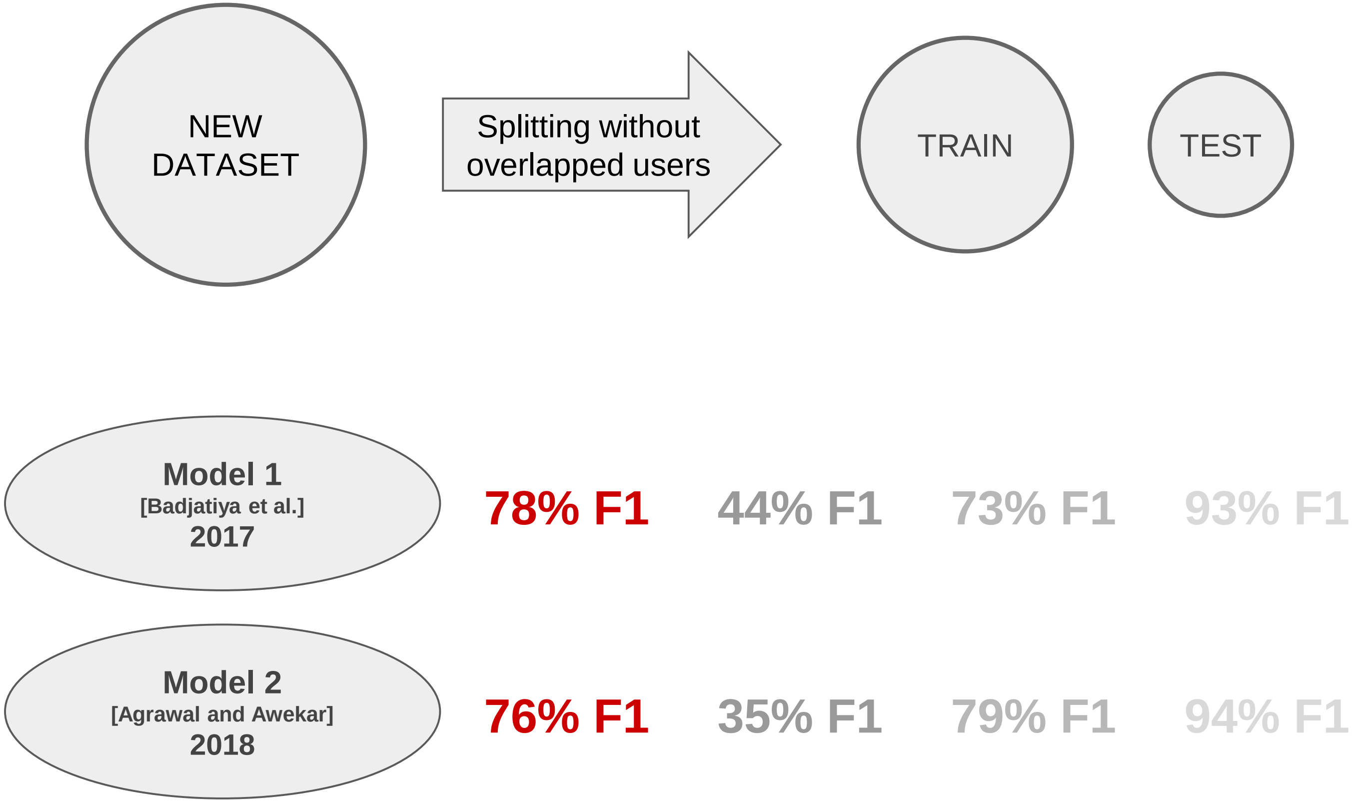
% Tweets from the most prolific user per class




DATASET 1
[Waseem and Hovy]
NAACL
2016
Splitting without overlapped users
TRAIN
TEST
Model 1
[Badjatiya et al.]
2017
44% F1
73% F1
93% F1
Model 2
[Agrawal and Awekar]
2018
35% F1
79% F1
94% F1

What happens if we have a dataset with a better user distribution?


NEW
DATASET
DATASET 1
250 tweets
per user
per class
DATASET 2
Hateful tweets


NEW DATASET
Splitting without overlapped users
TRAIN
TEST
Model 1
[Badjatiya et al.]
2017
78% F1
44% F1
73% F1
93% F1
Model 2
[Agrawal and Awekar]
2018
76% F1
35% F1
79% F1
94% F1

User distribution on datasets has an impact on the classification results.

State-of-the-art replication

User distribution

Generalization


TRAINING SET
TESTING SET

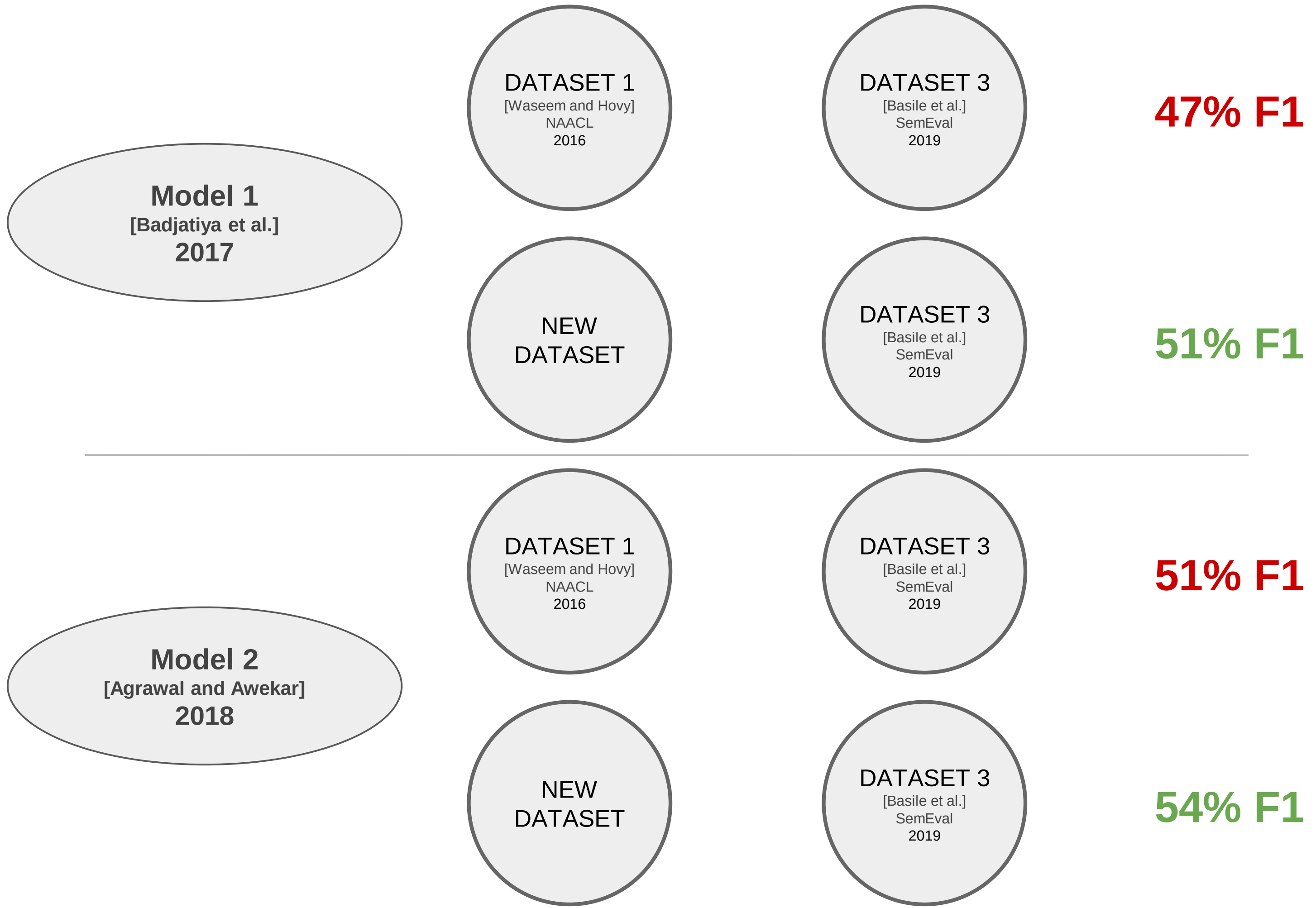
TRAINING SET

DATASET 3

[Basile et al.]

SemEval

2019


Model 1
[Badjatiya et al.]
2017
DATASET 1
[Waseem and Hovy]
NAACL
2016
DATASET 3
[Basile et al.]
SemEval
2019
47% F1
NEW DATASET
DATASET 3
[Basile et al.]
SemEval
2019
51% F1
Model 2
[Agrawal and Awekar]
2018
DATASET 1
[Waseem and Hovy]
NAACL
2016
DATASET 3
[Basile et al.]
SemEval
2019
51% F1
NEW DATASET
DATASET 3
[Basile et al.]
SemEval
2019
54% F1

Better user-distributed datasets lead to better generalization.

Conclusions

Hate Speech Detection is Not as Easy as You May Think

We show that state of the art results are highly overestimated due to experimental issues in the models:

Including the testing set during training phase

Oversampling the data before splitttting

User-biased datasets

Hate Speech Detection is Not as Easy as You May Think:

A Closer Look at Model Validation

Aymé Arango, Jorge Pérez and Bárbara Poblete

Universidad de Chile

Instituto Milenio Fundamentos de los datos