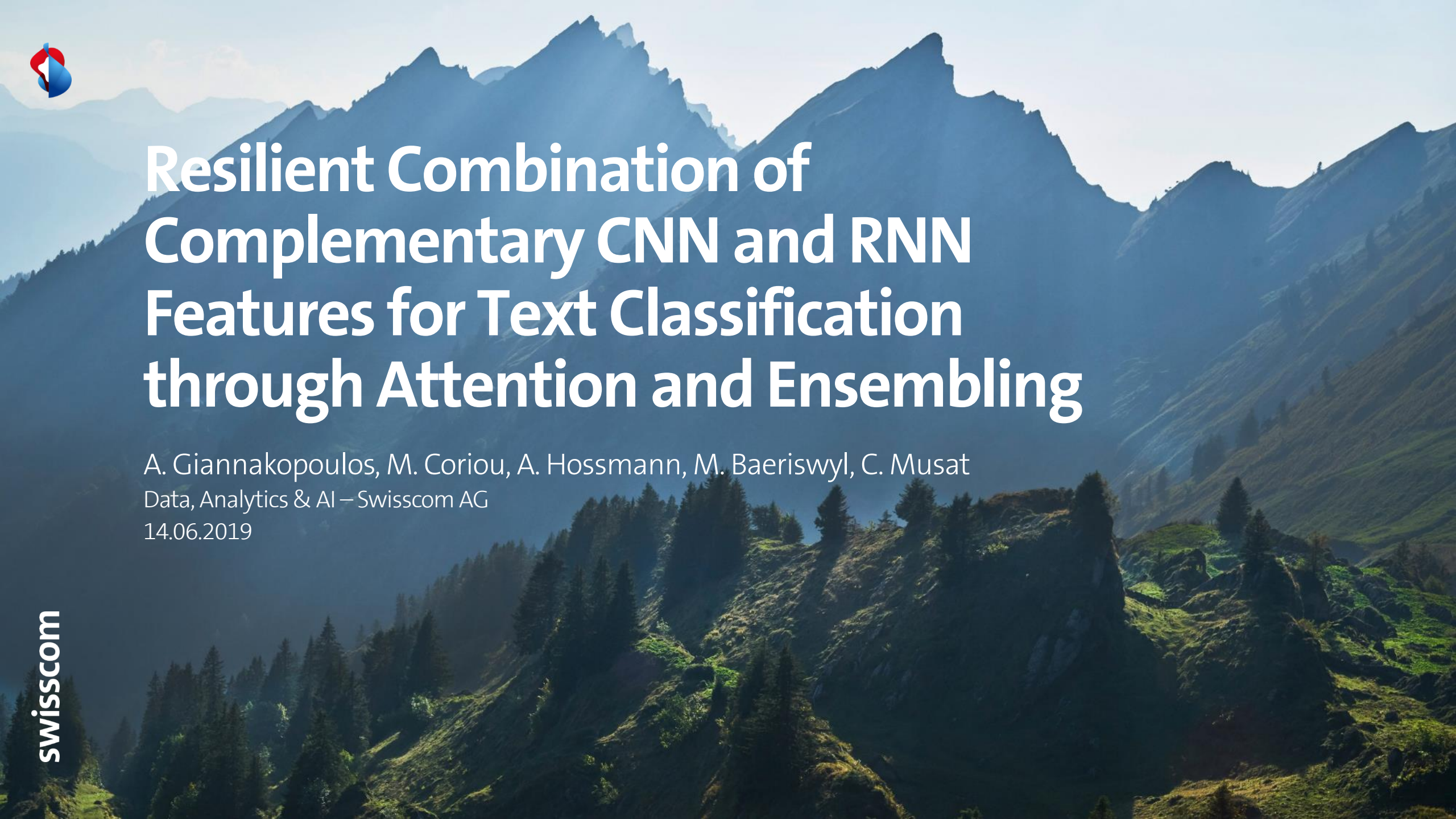


# Resilient Combination of Complementary CNN and RNN Features for Text Classification through Attention and Ensembling

A. Giannakopoulos, M. Coriou, A. Hossmann, M. Baeriswyl, C. Musat
Data, Analytics & AI – Swisscom AG
14.06.2019

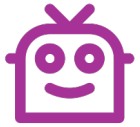
swisscom

# What drives us?

Based on Data, Analytics and AI we

Build new market offerings

Empower better decision-making



Enable Automation

## By enabling automation we

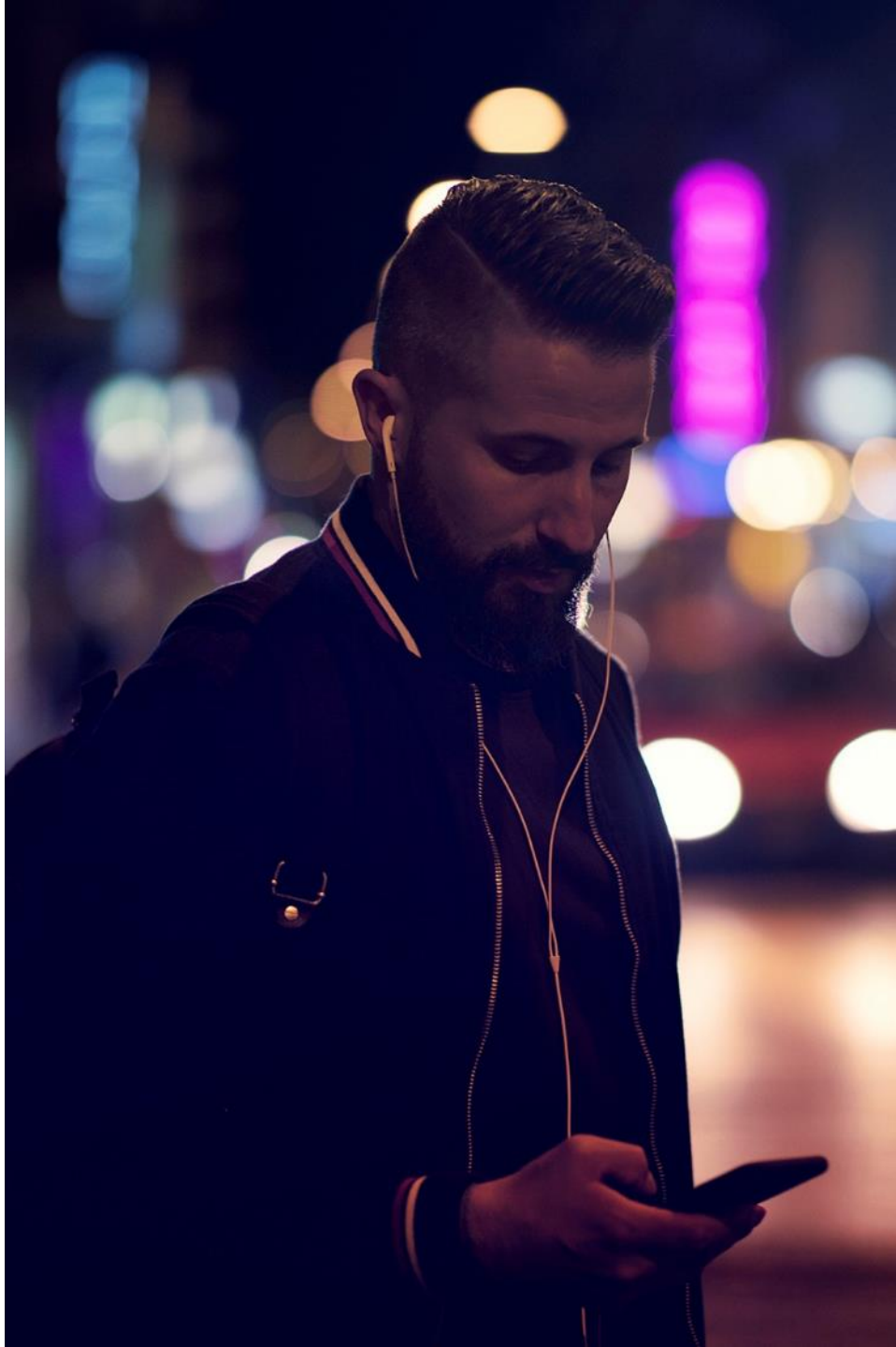
- Reduce operational cost
- Create faster responses to user requests
- Improve service quality of agents
- Improve user experience

## Text is the biggest source of logged data

Our main communication channels are

Emails

Chatbot



# Unsuitability of Existing Classifiers

State-of-the-art NLP models
(for research)



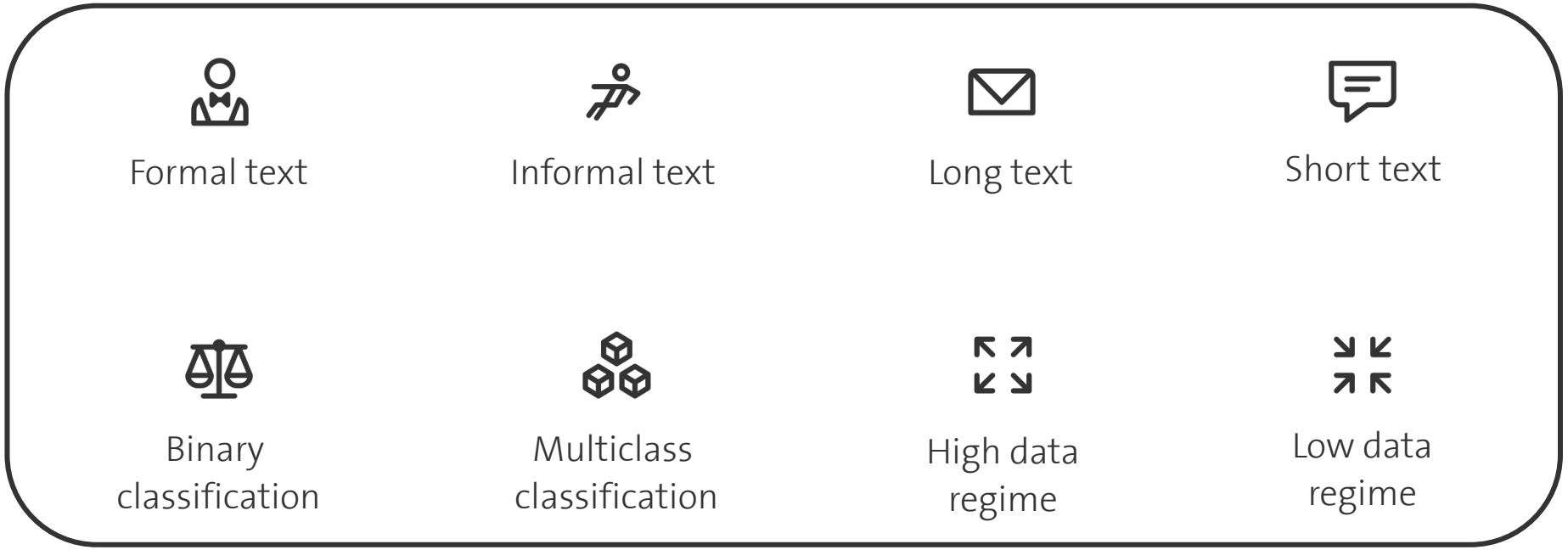

State-of-the-art NLP models
(for production)

- Difficult to fine-tune
- Difficult to deploy
- Difficult to scale

Complexity — trade-off — Performance

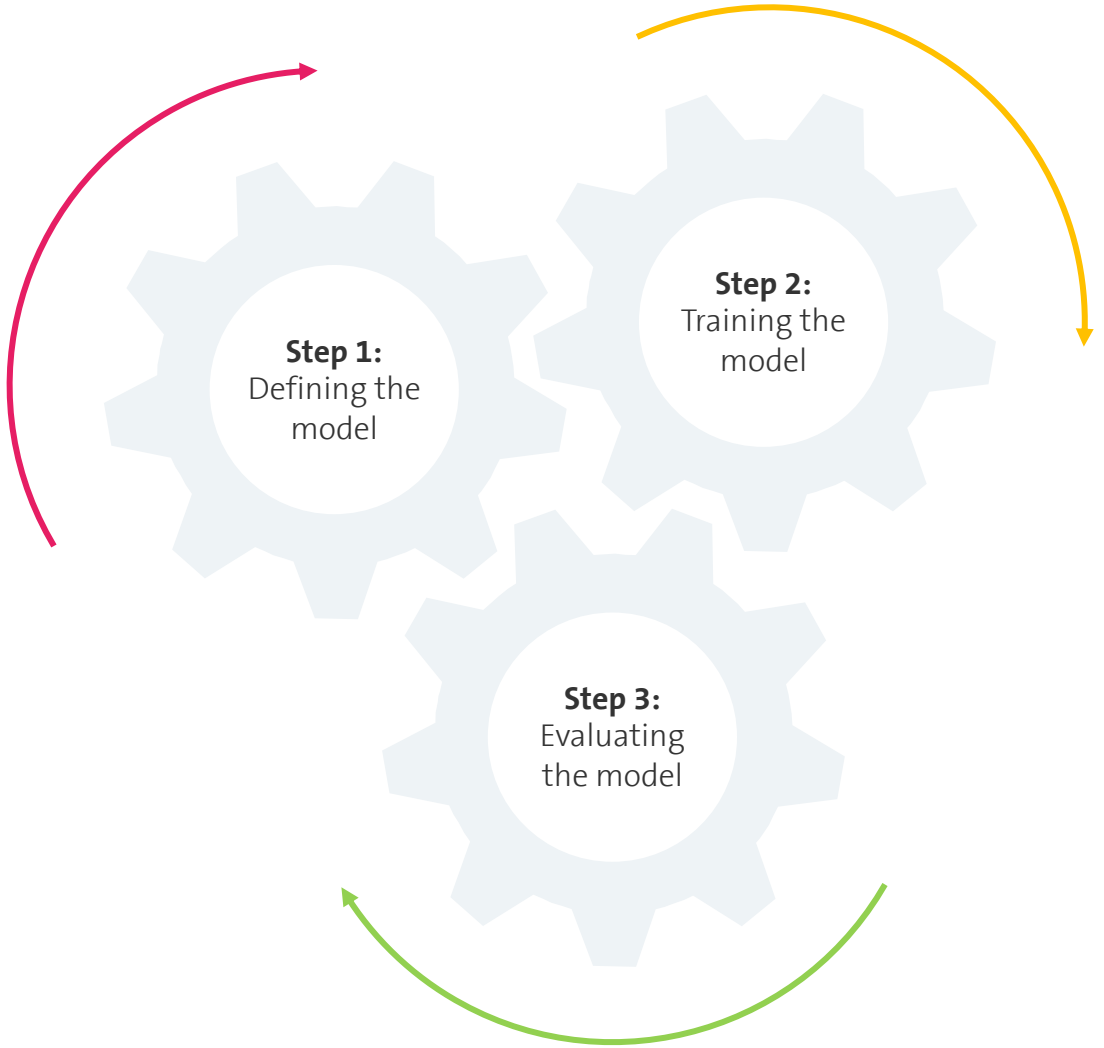
# Motivation for our Text Classifier

Can we build a text classifier for...

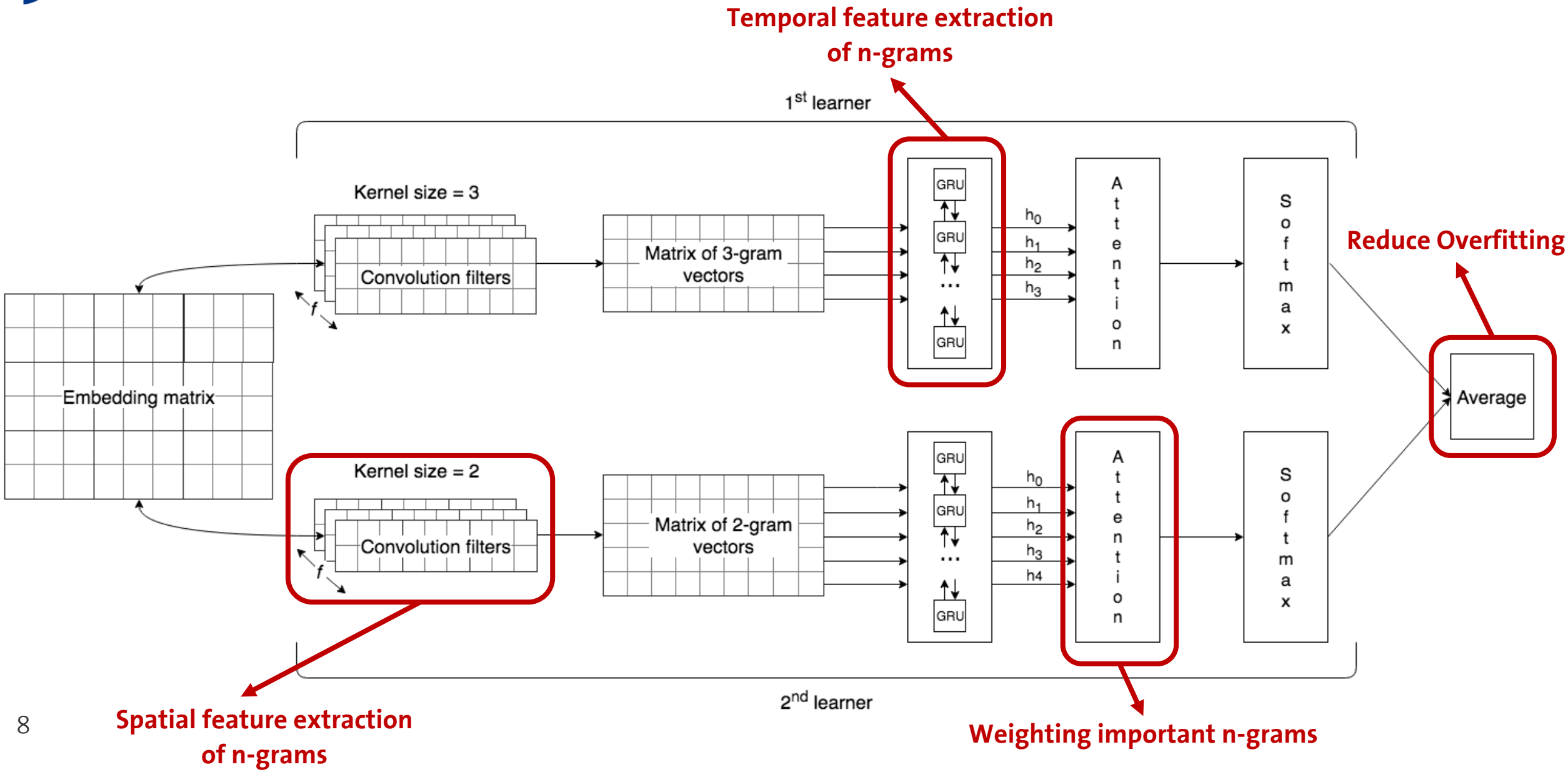



... and all of those at the same time?

# Machine Learning Pipeline



# ECGA Architecture: Ensemble of CNN, GRU and Attention



# Task Description

### Topic Classification

We use the DBpedia dataset

- 560 000 datapoints for training
- 70 000 datapoints for testing
- Balanced among 14 classes

- Long text
- Formal text
- Multiclass classification
- High data regime

### Argumentation Mining

We use the chilean constitution dataset

- More than 200 000 datapoints
- No official train/test split
- Up to 44 classes depending on the task

- Long text
- Formal text
- Multiclass classification
- High data regime

### Textual Churn Detection

We use a twitter dataset from 3 Telco providers in the USA

- Approx. 5600 tweets
- 10-fold CV for evaluation
- Imbalanced dataset (approx. 20% are churny tweets)

- Short text
- Informal text
- Binary classification
- Low data regime

# Dataset Preprocessing

| | Topic Classification (DBpedia) | Argumentation mining (Chilean constitution) | Churn detection (tweets) |
|---|---|---|---|
| Tokenization | ✗ | ✗ | ✗ |
| Stopword removal | ✗ | ✗ | |
| Lemmatization | ✗ | | |
| URLs, smileys, etc | | | ✗ |
| Padding | 60 | 60 | 50 |
| Vocabulary size | All tokens | All tokens | 1000 |
| Word Embeddings | fastText (EN) | fastText (ES) | GloVe (EN) |

# Experimental Setup

| Architecture Layer | Hyper-parameters |
|---|---|
| Convolution (CNN) | number of convolution filters |
| | size of 1D convolution |
| Recurrence (GRU) | number of units |
| Attention (ATT) | size of square matrix |
| Ensembling | number of learners |
| Dropout between all layers<br>Training is done with the *adam* optimizer<br>Early stopping with a patience of 10 epochs | |

# Reporting Results

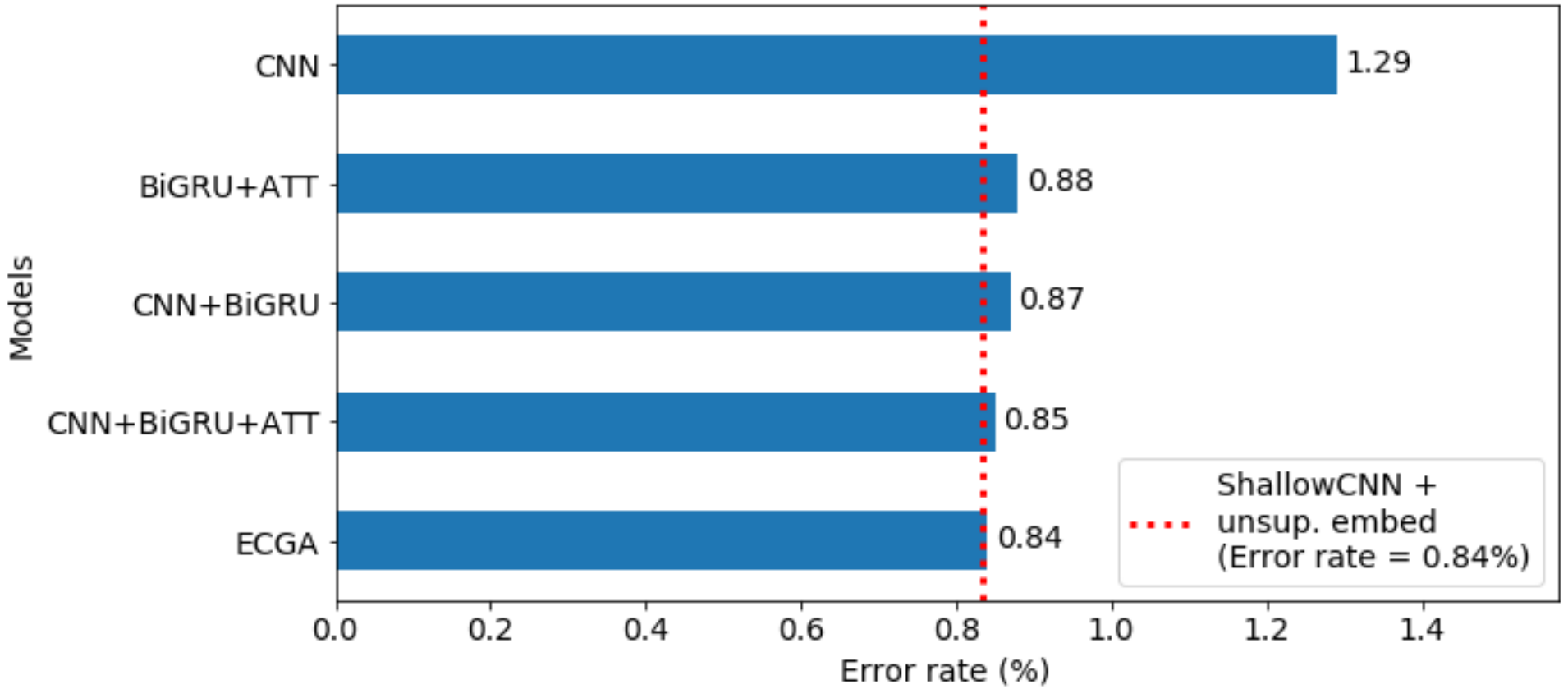
Hyper-parameter optimization using grid search

5 experiments for each hyper-parameter setup

Experiments last from 7 to 12 hours depending on the dataset

Average performance is reported

# Results on Topic Classification




More modules give higher performance

# Results on Argumentation Mining

| Models | Task A (Accuracy) | | | | | Task C (F1) |
|---|---|---|---|---|---|---|
| | Values<br>37 classes | Rights<br>44 classes | Duties<br>12 classes | Institutions<br>21 classes | Avg. | |
| Baseline<br>(Fierro et al., 2017) | 68.0 | 71.1 | 76.9 | 69.4 | 71.4 | 65.4 |
| CNN | 70.0 | 72.0 | 76.0 | 70.4 | 72.1 | 67.0 |
| BiGRU+ATT | 72.3 | 74.4 | 77.3 | 71.9 | 73.9 | 70.9 |
| CNN+BiGRU | 70.9 | 73.8 | 77.0 | 71.2 | 73.2 | 70.4 |
| CNN+BiGRU+ATT | 72.2 | 74.2 | 77.6 | 72.1 | 74.0 | 71.3 |
| ECGA | **72.5** | **75.0** | **78.2** | **72.4** | **74.5** | **71.6** |

Task A: SOTA with 4.3% relative improvement

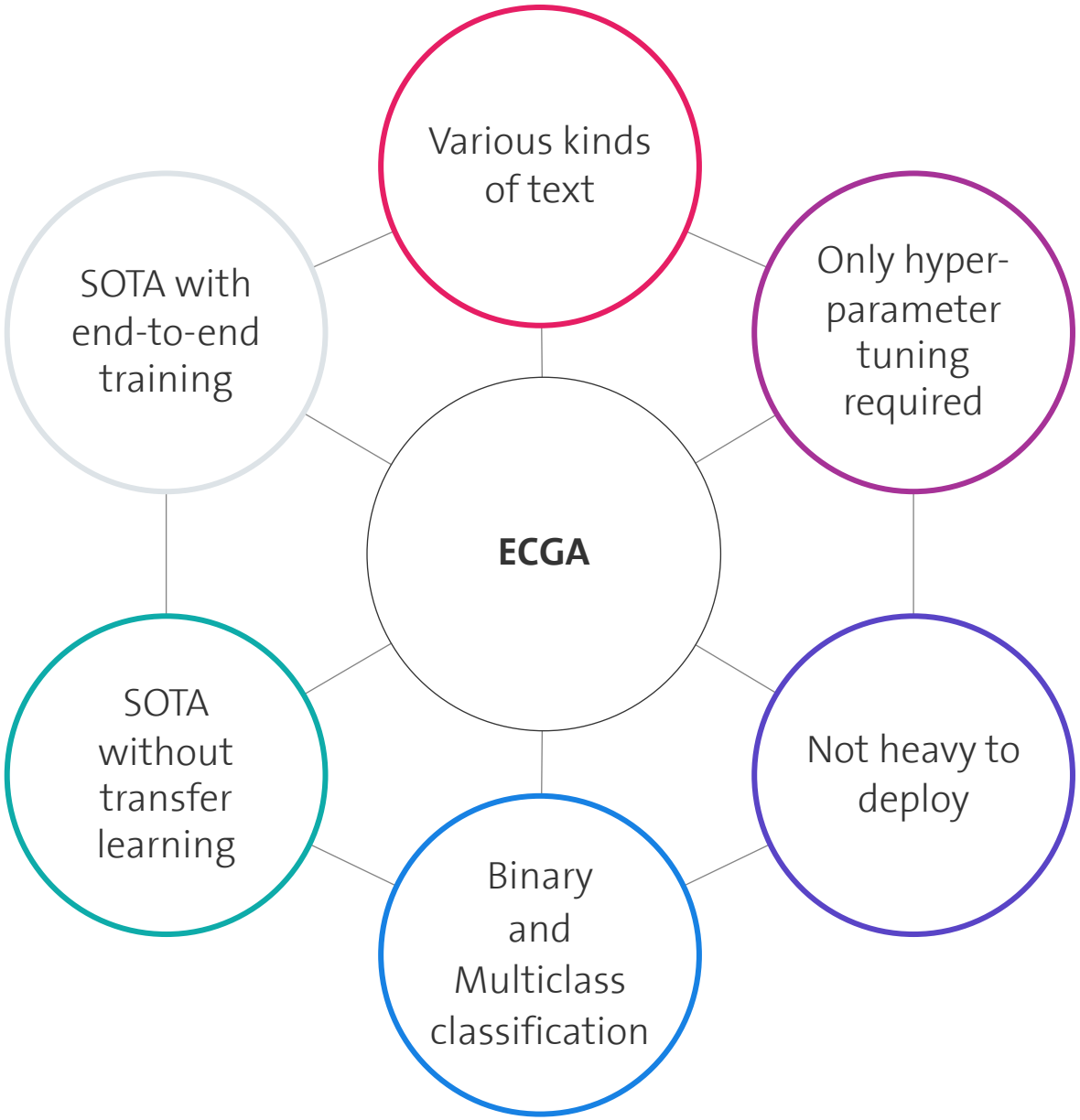
Task C: SOTA with 9.5% relative improvement

# Results on Churn Detection

| Models | AT & T | Verizon | T-Mobile | Macro F1 |
|---|---|---|---|---|
| CNN (Gridach et al, 2017) | | | | 80.67 |
| CNN + rules (Gridach et al, 2017) | | | | 83.85 |
| CNN | 78.71 | 81.52 | 85.11 | 81.94 |
| BiGRU+ATT | 82.14 | 84.15 | 85.85 | 84.21 |
| CNN+BiGRU | 83.52 | 83.93 | 85.19 | 84.48 |
| CNN+BiGRU+ATT | 85.02 | 85.73 | 87.36 | 86.26 |
| ECGA | 84.55 | 88.57 | 87.43 | **87.00** |

SOTA with 7.8% (and 3.7%) relative improvement on Macro F1

# Conclusion




- ✓ ECGA is a go-to text classifier that **achieves or surpasses SOTA** results in a variety of datasets
- ✓ Neural modules are complementary and **stacking** them **increases performance**
- ✓ ECGA is trained in an **end-to-end** way
- ✓ Training time is **acceptable** (avg. 14 hours on a GPU server)