Challenges in Applying Machine Learning Methods: Studying Political Interactions on Social Networks

CHAYA LIEBESKIND* AND KARINE NAHON~

*Jerusalem College of Technology, Lev Academic Center, Jerusalem/Israel
~ Interdisciplinary Center Herzliya, Israel and University of Washington, USA

Social Networks

A vast amounts of user-generated content

An opportunity for research to understand behavioral questions

Political Interactions

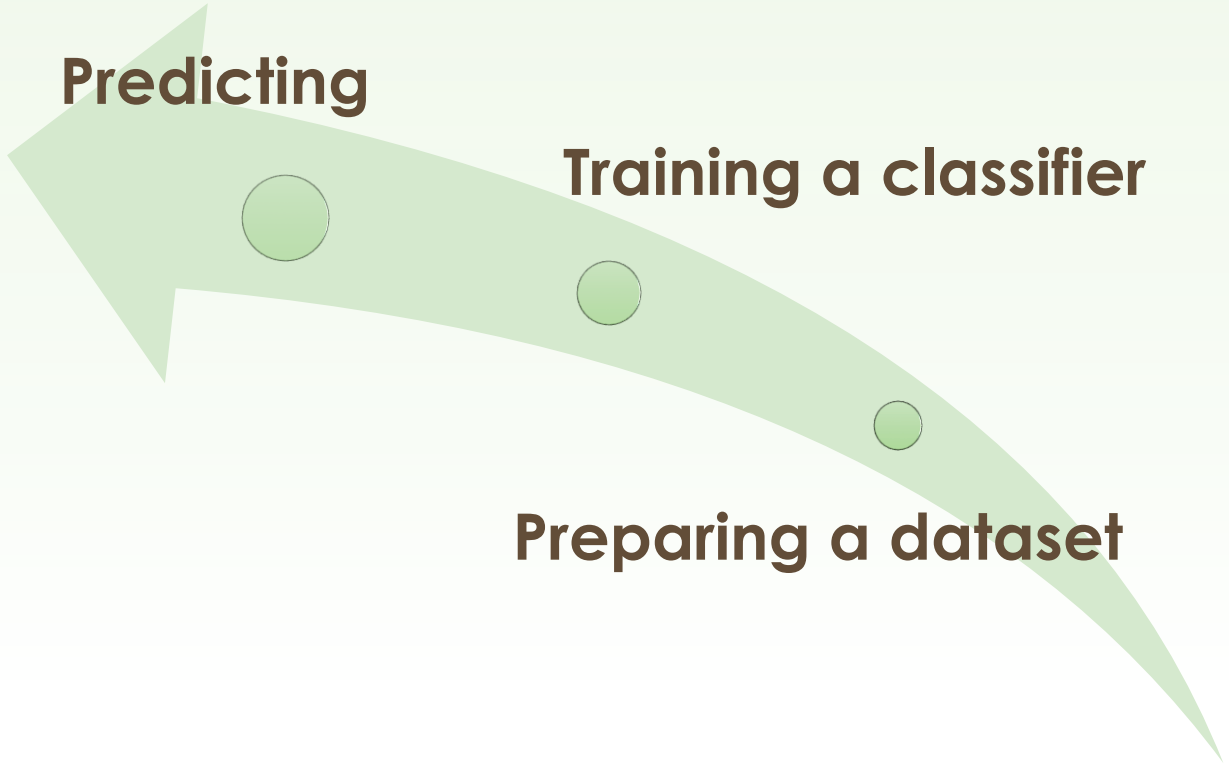
Machine Learning

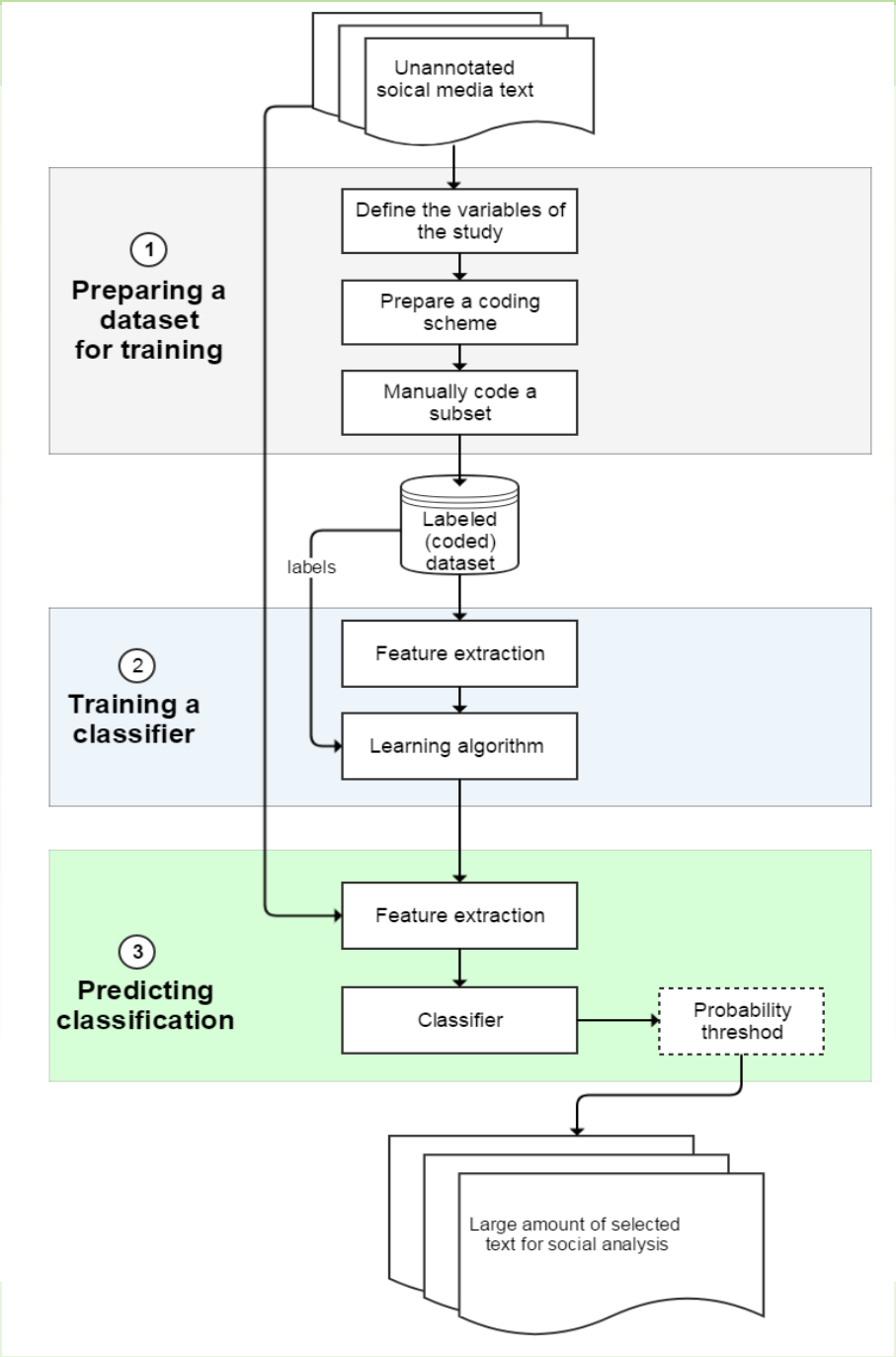
- Manual content analysis
  - Requires high levels of efforts and time to code and analyze

- Machine Learning
  - Analyze vast amounts of data automatically

Supervised machine learning methods for political-orientated classification tasks

Supervised Machine Learning

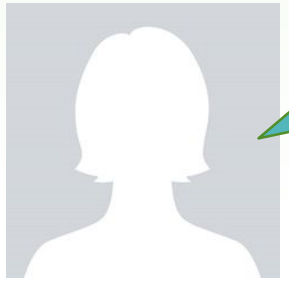




Challenges in classifying relevance of political comments while using supervised ML techniques

Comment Relevance Classification

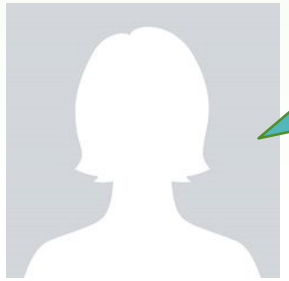
"I am speaking now about the security situation in Israel. I will address the lies that the Palestinian Authority continues to tell."



"This is the truth sayings by Prime Minister of Israel..."

Comment Relevance Classification

"The danger in the coming elections is the establishment of a leftist government..."



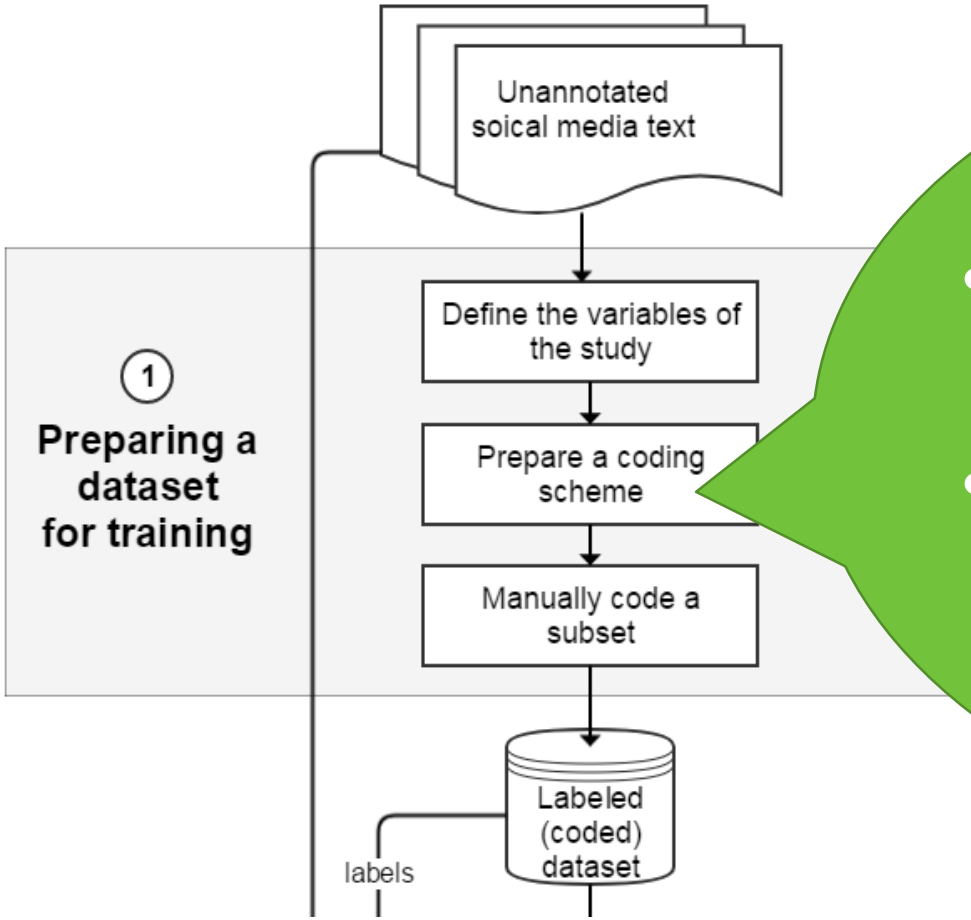
"Would love to have seen this sub-titled in English!"

Comment Relevance Classification in Facebook

- A corpus of 4.8 million comments written in Hebrew by users replying to 41,882 politicians' posts
  - Posted on Facebook during 2014-2015
  - Average length of a comment is 7 words
  - Average length of a post is 22 words

- A sub-corpus of 1,397 comments was manually annotated for relevance classification
  - 803 positive examples and 594 negative examples

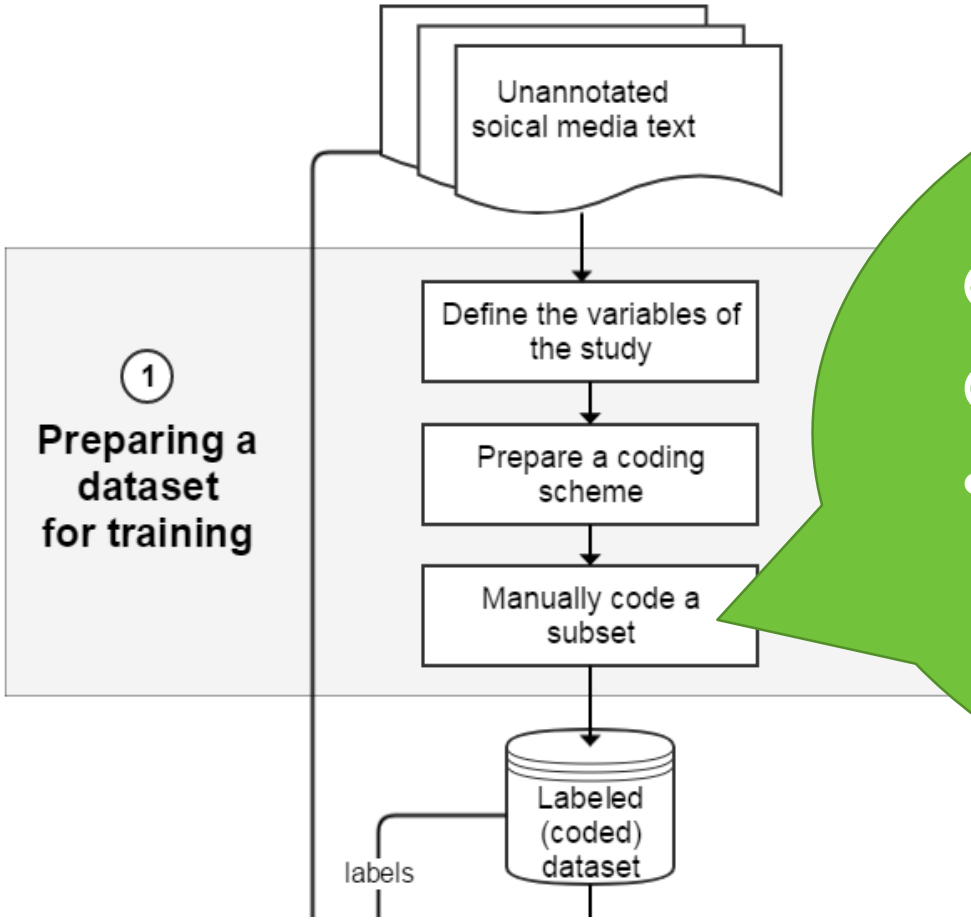
Preparing a dataset for training



An iterative process:

- Requires further refinement of the coding guidelines
- Until reaching an appropriate inter-rater reliability of agreement

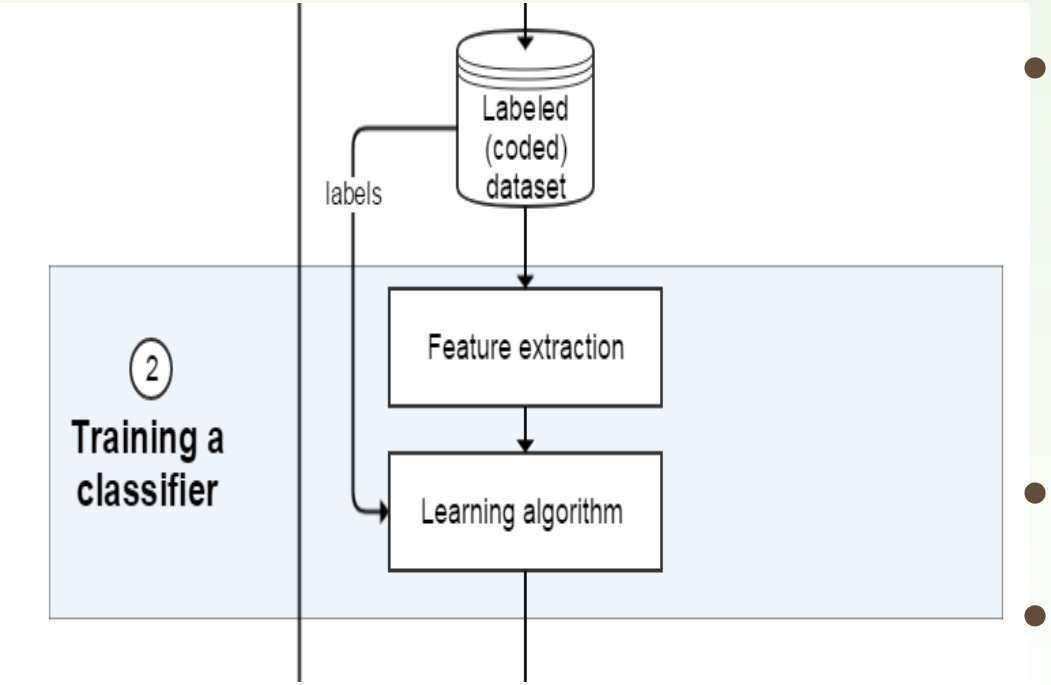
Preparing a dataset for training



The subset of training examples should follow the distribution of the data

- Under-sampling MKs on the 'long tail'

Training a classifier



- Extracting a feature set
  - Word representation
  - Character n-grams representation
  - Metadata features
- Applying feature selection methods
- Enriching the feature set to optimize the classification performance

Training a classifier

A comparison of character n-grams configurations

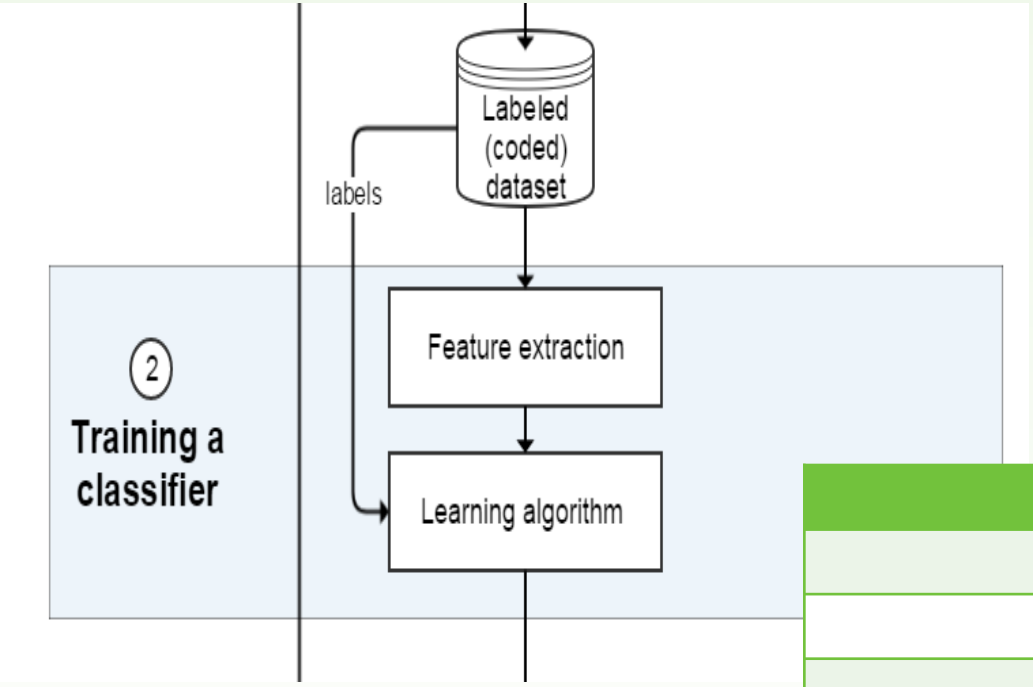
Input is the comment text:

Character N-grams	Accuracy (%)	F-Measure
n=2	63.72	0.75
n=3	69.23	0.78
n=4	68.48	0.77
n=5	**69.57**	**0.78**

Input is both the post and the comment text:

Character N-grams	Accuracy (%)	F-Measure
n=2	68.14	0.74
n=3	59.7	0.78
n=4	**76.79**	**0.82**
n=5	72.9	0.8

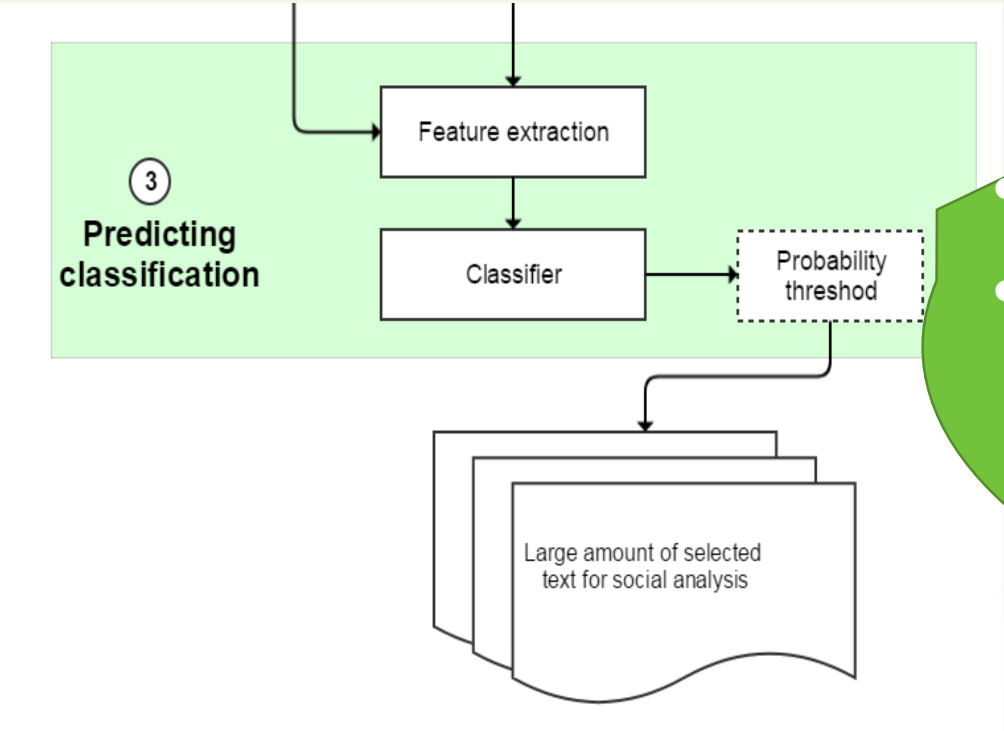
Training a classifier



- Selecting a supervised learning algorithm
- Analyzing the classification results

ML method	Accuracy %	F-Measure
RandomForest	73.52	0.78
Decision Tree	63.1	0.72
Bayes Network	59.9	0.72
Supported Vector Machine (SVM)	76.79	0.82
Logistic Regression	79.17	0.83
Bagging	71	0.77
AdaBoost	60.11	0.73

Predicting classification of big data



- To achieve a higher accuracy
- Use algorithms that produce probabilities of membership (P(class | input))

we are currently running our trained classifier to predict the comment relevance classification of over than 5M comments


Thank You