NLP and Text Sentiment Analysis summary

A. Zamichos, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr

Version 1.1.1

Artificial Intelligence & Information Analysis Lab

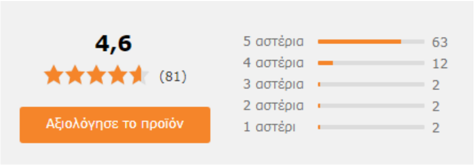
NLP and Text Sentiment Analysis

- Baseline algorithms
- Text pre-processing
- Word embeddings
    - Word2Vec
    - Fast Text
- NLP and Sentiment Analysis - A text classification task
    - Neural Networks (RNN, CNN)
    - 2018-2019 NLP's ImageNet moment (Transfer Learning)
        - Contextual Embeddings (ELMo)
        - BERT

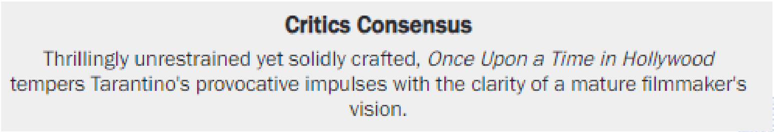
What is Sentiment Analysis

Examples of sentiment analysis

- **Products**: what do people think about new iPhone ?



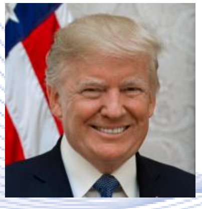
- **Movies**: what do people think about DiCaprio’s new movie ?



Critics Consensus

Thrillingly unrestrained yet solidly crafted, *Once Upon a Time in Hollywood* tempers Tarantino's provocative impulses with the clarity of a mature filmmaker's vision.

- **Politics**: what do people think about a candidate (e.g. Trump) or an issue ?



 or  ?

- **Emotional state** (advanced)**:** what is the emotional state of general public for an issue such as (e.g. angry, sad, happy)

Why Sentiment Analysis

Artificial Intelligence &
Information Analysis Lab

Why sentiment analysis?

- Allows companies to make sense of the available unstructured text
  - 80% of the world’s data is unstructured
- Saves hours of manual data processing
- Real time analysis
- Scalability

Sentiment analysis other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

A Text Classification Task

Sentiment Analysis

Three levels of granularity

- Sentence level
- Paragraph level
- Document level

Baseline algorithms

Sentiment Analysis

Rule-based methods

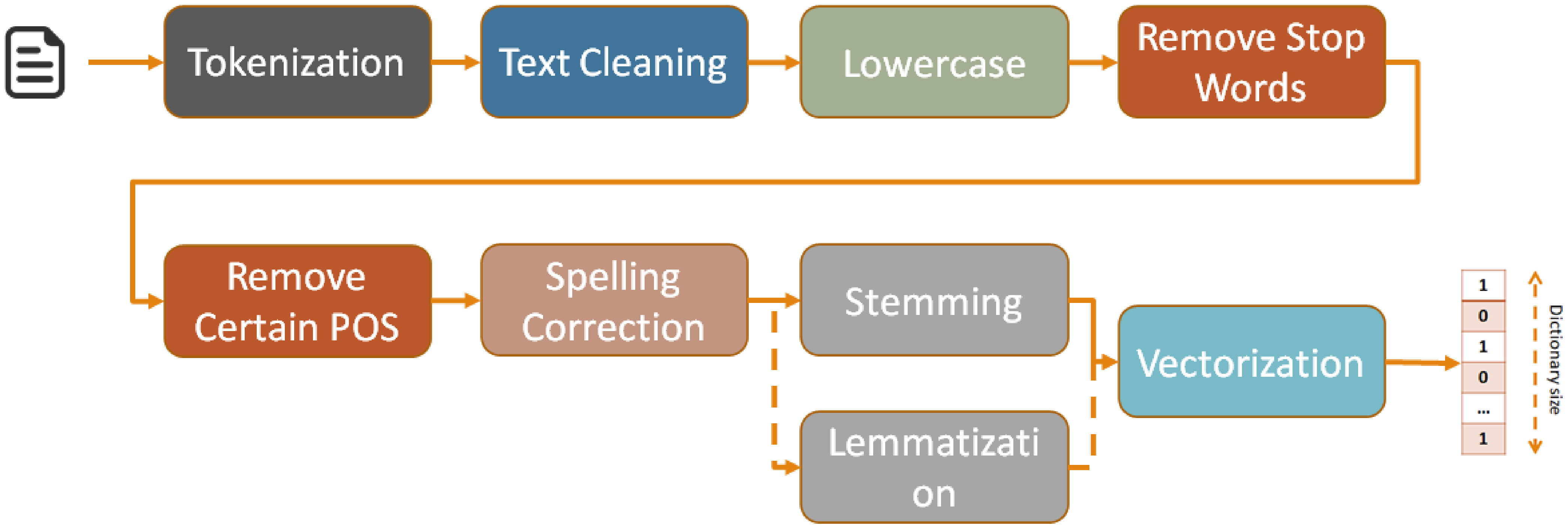
- Rule-based methods (use of lexicons)
  - Count the positive and the negative words within text
  - Lexicons with sentiment of the words
  - Some available lexicons:
    - https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html#lexicon (6800 positive and negative words for English language)
    - https://www.kaggle.com/rtatman/sentiment-lexicons-for-81-languages#correctedMetadata.csv (sentiment lexicons for 81 languages)
    - https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/ WordStat Sentiment Dictionary: includes more than 9164 negative and 4847 positive word patterns
    - SentiWordNet: a lexical resource -> assigns to each synset of WordNet three sentiment scores: positivity, negativity, and objectivity

Supervised methods (examples)

- Naïve Bayes Classifier
- K-nearest Neighbor
- Support Vector Machine (SVM)
- Decision Tree
- Random Forest
- Boosting algorithms
- Bagging algorithms

Text Pre-processing

Common pre-processing steps



Word embeddings

Word Embeddings

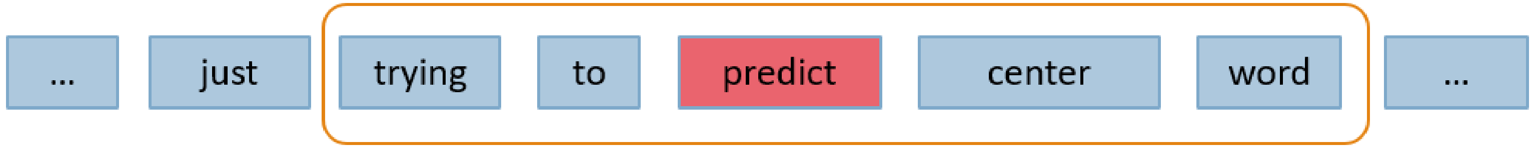
- Key breakthrough for NLP
- Words are represented as real-valued vectors
- Words with same meaning have similar representations
- Most common methods
  - **Word2Vec** (T. Mikolov et al, 2013) - Google
  - **Glove** (J. Pennington et al, 2014) - Stanford
  - **FastText** (P. Bojanowski et al, 2016) – Facebook AI Research
- word embedding for the word “king” (GloVe vector trained on Wikipedia)

```
[ 0.50451 , 0.68607 , -0.59517 , -0.022801, 0.60046 , -0.13498 , -0.08813 , 0.47377 , -0.61798 , -0.31012 ,
-0.076666, 1.493 , -0.034189, -0.98173 , 0.68229 , 0.81722 , -0.51874 , -0.31503 , -0.55809 , 0.66421 , 0.1961
, -0.13495 , -0.11476 , -0.30344 , 0.41177 , -2.223 , -1.0756 , -1.0783 , -0.34354 , 0.33505 , 1.9927 ,
-0.04234 , -0.64319 , 0.71125 , 0.49159 , 0.16754 , 0.34344 , -0.25663 , -0.8523 , 0.1661 , 0.40102 , 1.1685 ,
-1.0137 , -0.21585 , -0.15155 , 0.78321 , -0.91241 , -1.6106 , -0.64426 , -0.51042 ]
```

Word2Vec (concept)

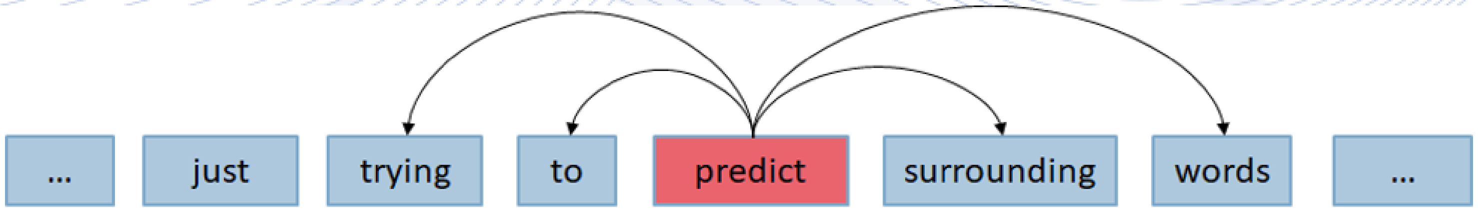
- Continuous Bag-of-Words (CBOW)
  - Predict center word given surrounding words

input1	input2	input3	input4	output
trying	to	center	word	predict



- Skip-Gram
  - Predict surrounding words given center word

Input	output1	output2	output3	output4
predict	trying	to	center	word



Fast Text

- Scope
  - solve out of vocabulary problem (OOV) – word2vec, Glove
  - Better representation for languages with lots of morphologic

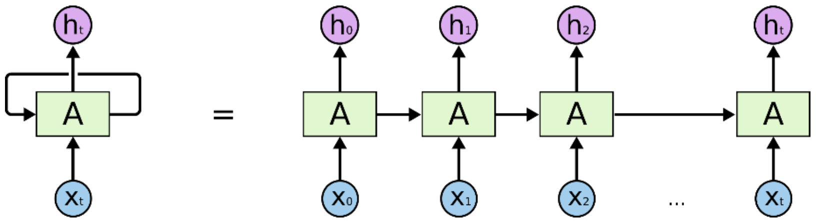
- Idea
  - Represent word as character n-grams
  - An extension of word2vev skip-gram model
  - E.g. where = <wh, whe, her, ere, re>, <where>
  - Represent word as sum of these representations

Neural Networks

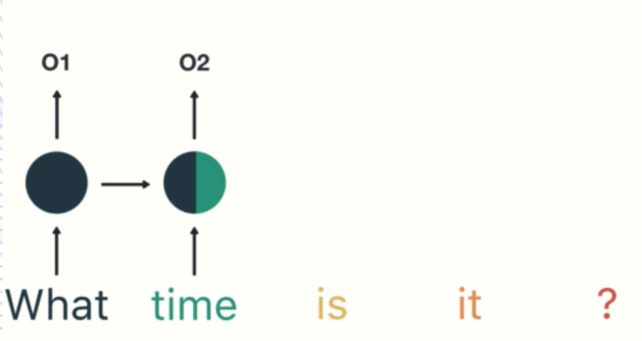
(RNN, CNN)

Recurrent Neural Networks (RNNs)

- Good for dealing with sequential data (as text)
- Consider information of previous nodes
- Why is it useful ?
  - Example: Try to predict the direction of a ball moving
- RNNs mostly works by using LSTM or GRU for text classification (due to vanishing gradient problem)



Source: colah's blog

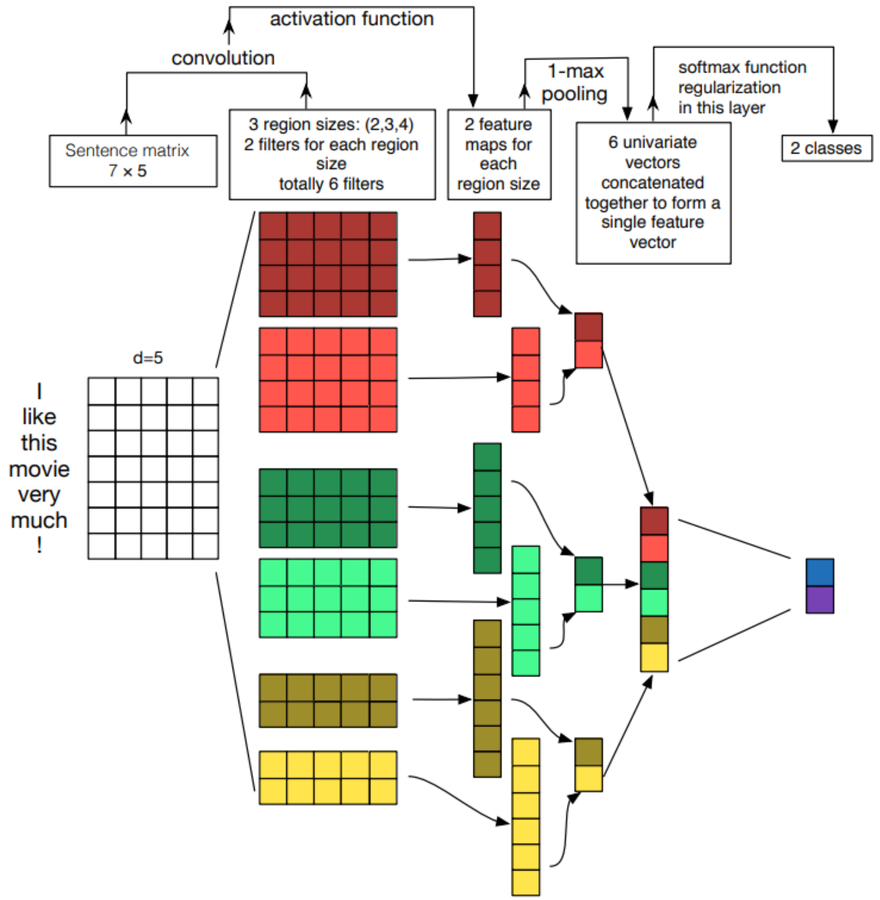


Source: link

Artificial Intelligence & Information Analysis Lab

Convolution Neural Networks (CNNs)

- Traditionally used in image processing
- Can be also used for text
- Easily parallelized for GPUs
- Good classification results



Source: Zhang and Wallace (2015)

2018-2019 NLP’s ImageNet moment

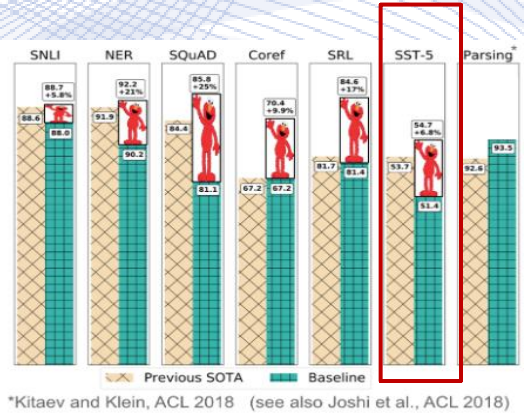
Transfer Learning

Contextual Embeddings (ELMo)

ELMo

- Matthew E. Peters et al (2018): Deep contextualized word representations
  - Idea:
    - Why to give predefined embeddings for words (word2vec, glove, fastText) ?
    - Same words have different meaning according to the sentence they used in.
    - uses a bi-directional LSTM trained to predict the next word in a sequence of words in a huge corpus of text
  - A significant step towards pre-training in the context of NLP
  - Achieved improvements on a wide range of NLP tasks.

Sources: Jay Alammar's -The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)



Sentiment analysis task

Artificial Intelligence & Information Analysis Lab

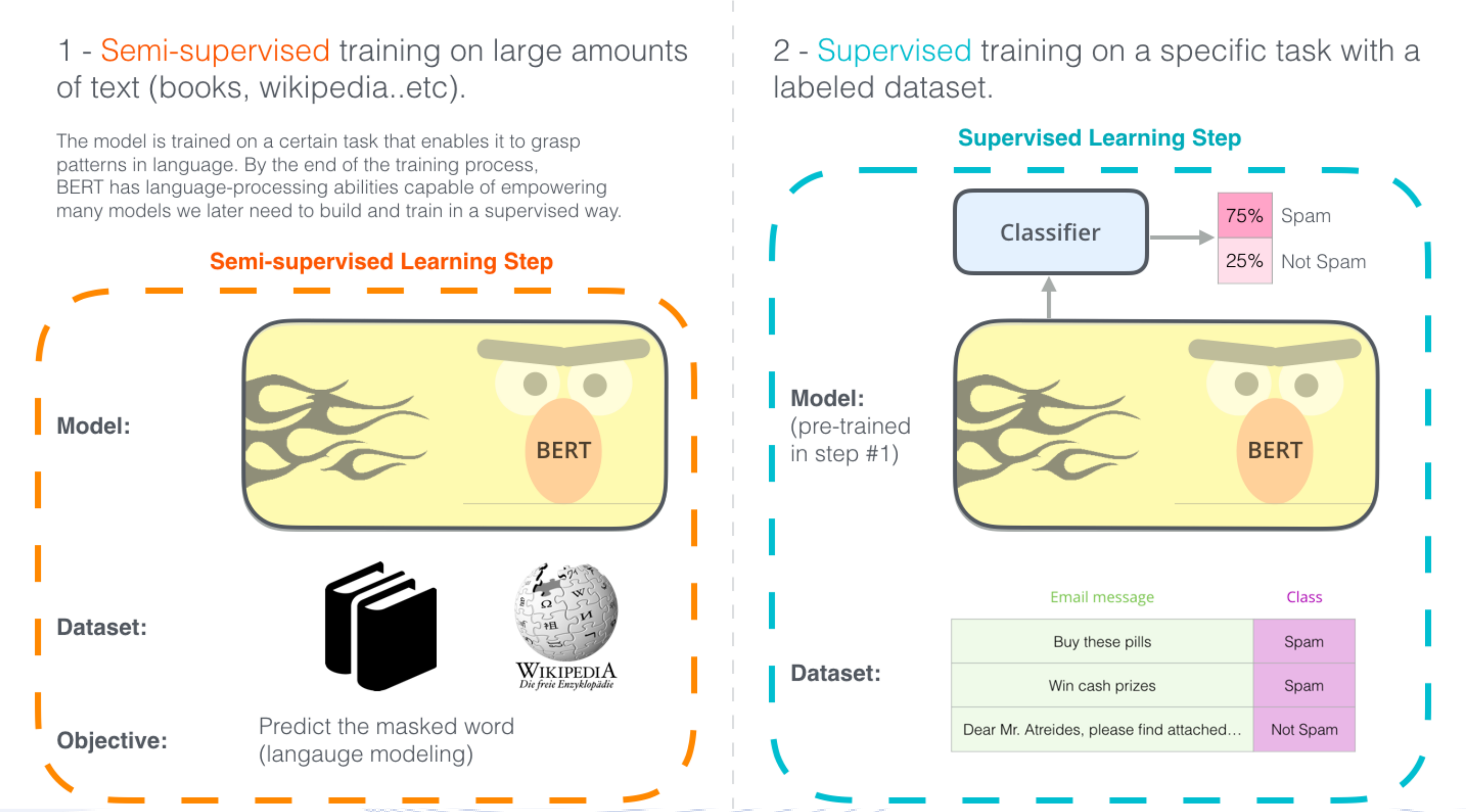
BERT

BERT

- Jacob Devlin et al (2019): BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

- builds on top of a number of clever ideas *
  - including but not limited to Semi-supervised Sequence Learning ,
  - ELMo ,
  - ULMFiT ,
  - the OpenAI transformer ,
  - and the Transformer (Vaswani et al) (Attention is all you need)

BERT



- Download the model pre-trained in step 1 (trained on un-annotated data),
- Fine-tuning it for step 2

Bibliography

[SAL2016] T. Salimans, D. P. Kingma "Weight normalization: A simple reparameterization to accelerate training of deep neural networks." *Advances in Neural Information Processing Systems*. 2016.

[MII2019] Miikkulainen, Risto, et al. "Evolving deep neural networks." *Artificial Intelligence in the Age of Neural Networks and Brain Computing*. Academic Press, 2019. 293-312.

Bibliography

[OPP2013] A. Oppenheim, A. Willsky, Signals and Systems, Pearson New International, 2013.

[MIT1997] S. K. Mitra, Digital Signal Processing, McGraw-Hill, 1997.

[OPP1999] A.V. Oppenheim, Discrete-time signal processing, Pearson Education India, 1999.

[HAY2007] S. Haykin, B. Van Veen, Signals and systems, John Wiley, 2007.

[LAT2005] B. P. Lathi, Linear Systems and Signals, Oxford University Press, 2005.

[HWE2013] H. Hwei. Schaum's Outline of Signals and Systems, McGraw-Hill, 2013.

[MCC2003] J. McClellan, R. W. Schafer, and M. A. Yoder, Signal Processing, Pearson Education Prentice Hall, 2003.

Bibliography

[PHI2008] C. L. Phillips, J. M. Parr, and E. A. Riskin, Signals, Systems, and Transforms, Pearson Education, 2008.

[PRO2007] J.G. Proakis, D.G. Manolakis, Digital signal processing. PHI Publication, 2007.

[DUT2009] T. Dutoit and F. Marques, Applied Signal Processing. A MATLAB-Based Proof of Concept. New York, N.Y.: Springer, 2009

Bibliography

[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.
[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr