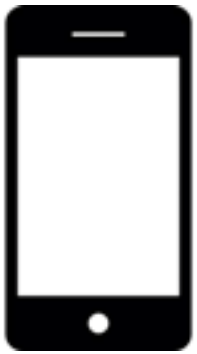
Analiza wydźwięku w krótkich wypowiedziach użytkowników[1]

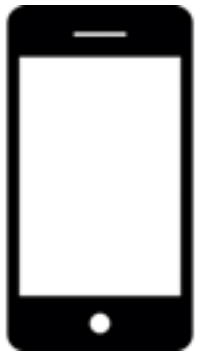
Mateusz Lango

21 czerwca 2016

[1] Lango M., Brzeziński D., Stefanowski J.: PUT at SemEval-2016 Task 4: The ABC of Twitter Sentiment Analysis, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), NAACL HLT 2016, San Diego, US

Why Sentiment Analysis?



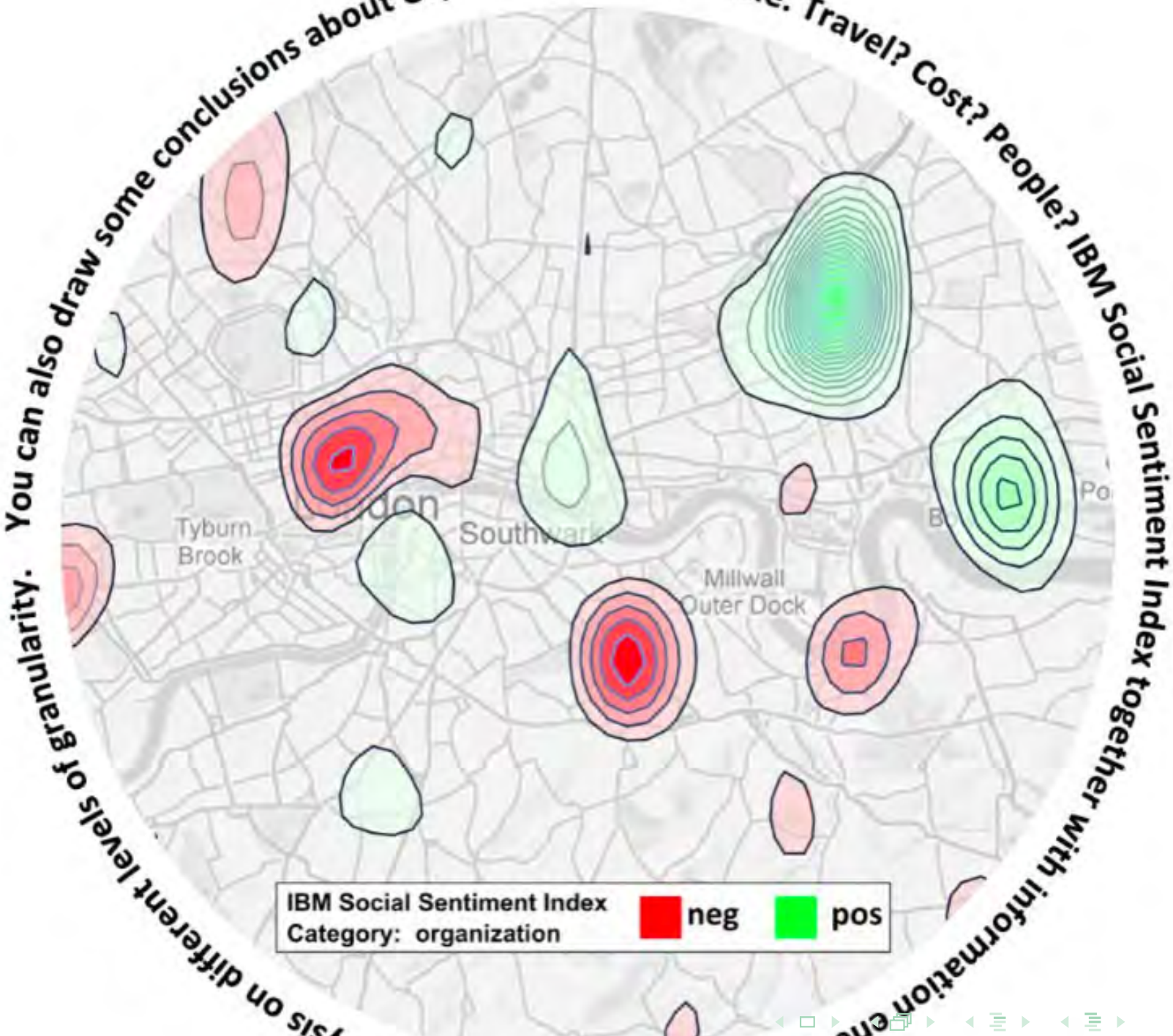


Weight	110 g	133 g
Resolution	480 x 640	320 x 480
RAM	256	128
HSDPA [Mbit/s]	7.2	3.6
Video call	Yes	No
Video recording	Yes	No
Voice commands	Yes	No
Voice recording	Yes	No
MMS	Yes	No
Memory cards	Yes	No

Why Sentiment Analysis?

	HTC Touch Diamond	Apple iPhone 3G
Weight	110 g	133 g
Resolution	480 x 640	320 x 480
RAM	256	128
HSDPA [Mbit/s]	7.2	3.6
Video call	Yes	No
Video recording	Yes	No
Voice commands	Yes	No
Voice recording	Yes	No
MMS	Yes	No
Memory cards	Yes	No

Why Sentiment Analysis?


You can also draw some conclusions about Olympics in macro-scale. Travel? Cost? People? IBM Social Sentiment Index together with information
Tyburn
Brook
Southwark
Millwall
Outer Dock
IBM Social Sentiment Index
Category: organization
neg
pos

“Energy of the Nation” project

Why Sentiment Analysis?

- Decision Support
- Product Design
- Market Research
- Social Science
- Machine Learning/Text Mining
- ...

Sentiment Analysis

- Document Sentiment Classification
- Sentence Subjectivity and Sentiment Classification
- Aspect-based Sentiment Analysis

Example of classical unsupervised approach

Pointwise mutual information

$$PMI(term_1, term_2) = \log_2 \frac{P(term_1, term_2)}{P(term_1)P(term_2)}$$

Sentiment orientation

$$SO(phrase) = PMI(phrase, excellent) - PMI(phrase, poor)$$

Sentiment Classification: Task Definition

- Input: An opinionated text object
- Output: A sentiment tag/label

Text preprocessing

- tokenization (!)
- lemmatization (!)
- stop-words removal (!)
- grouping rare, special tokens (urls, hashtags, numbers, percentages, prices, dates, hours)
- removing rare tokens (< 5)

N-grams

- word n-grams
- character k-grams
- POS n-grams
- elongated words
- emoticons
- punctuation
- all-caps

Negation problem

Review of "1Q84" by Haruki Murakami

Perhaps one of the most important works of science fiction of the year ... 1Q84 does not **disappoint** ... [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and entrancing. –Matt Staggs, Suvudu.com

Negation problem

Review of "1Q84" by Haruki Murakami

Perhaps one of the most important works of science fiction of the year ... 1Q84 **does not disappoint** ... [It] envelops the reader in a shifting world of strange cults and peculiar characters that is surreal and entrancing. –Matt Staggs, Suvudu.com

Negation n-grams

- negation list: not, never, none, nobody, nowhere, neither
- negation context from the word following the negation word until the next punctuation mark

The voice quality of this phone is not good, but the battery life is long

The room was very nicely appointed and the bed was sooo comfortable. Even though the bathroom door did not close all the way, it was still pretty private.

Sentiment Lexicons

- SentiWordNet
- Opinion Lexicon
- Multi-perspective Question Answering (4 categories)
- NRC (8 emotions)
- ...

How to annotate?

MaxDiff methodology

great good bad interesting

How to annotate?

MaxDiff methodology

great good bad interesting

How to annotate?

MaxDiff methodology

great good bad interesting

	great	good	bad	interesting
great		>	>	>
good				
bad	<	<		<
interesting				

How to annotate?

MaxDiff methodology

great good bad interesting

	great	good	bad	interesting
great	–	>	>	>
good	<	–	>	
bad	<	<	–	<
interesting	<		>	–

Hashtag Sentiment Lexicon

- assume that tweets with specific hashtags have known sentiment (e.g. #joy, #sad, #angry, #surprised)
- crawl tweets during 8 months
- filter very short&misspelled tweets
- use PMI
- investigate influence of negation
  - great [highly positive] $\rightarrow$ not great [mildly negative]
  - terrible [strong negative] $\rightarrow$ not terrible [mildly negative]

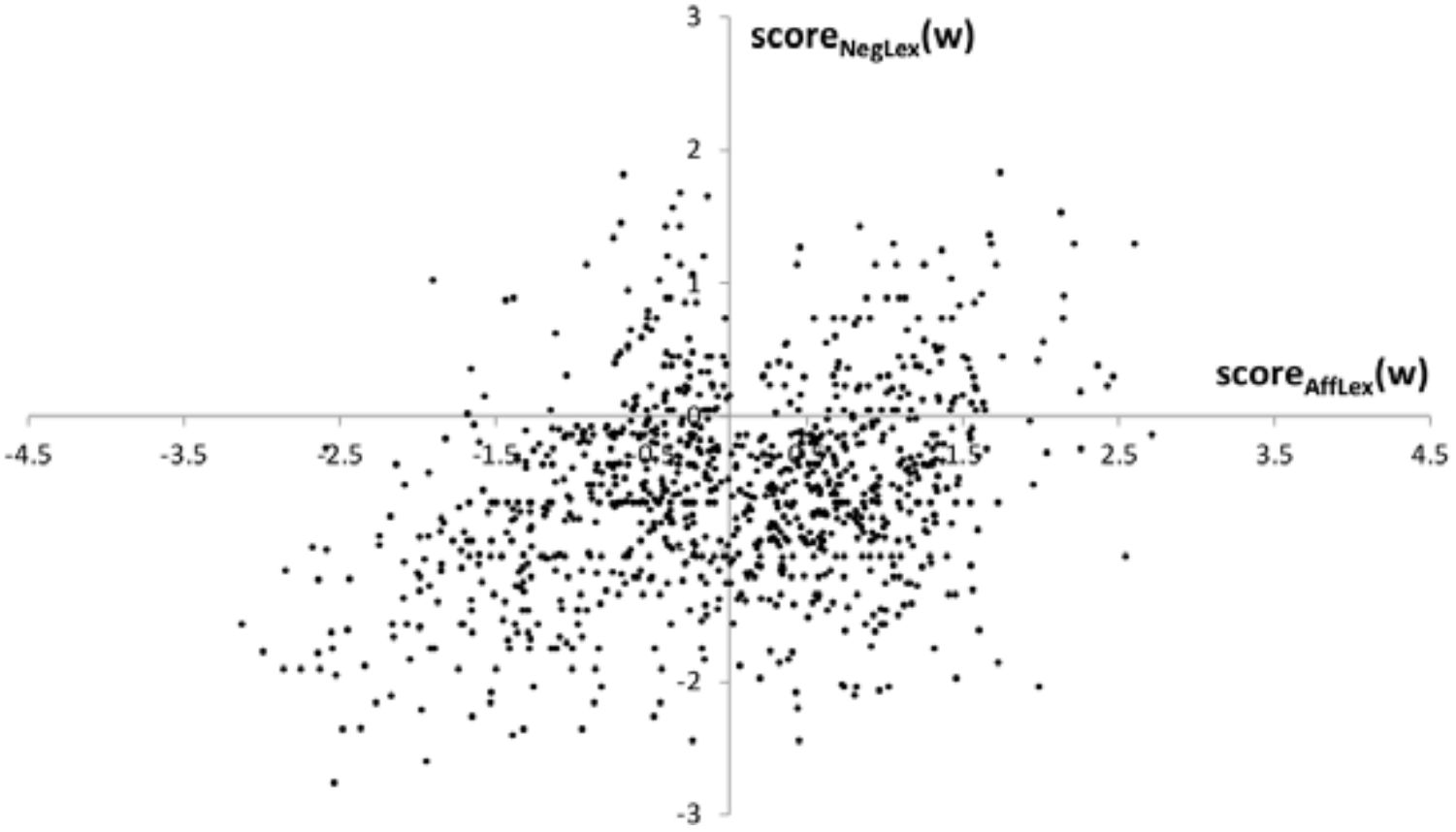
Hashtag Sentiment Lexicon

- assume that tweets with specific hashtags have known sentiment (e.g. #joy, #sad, #angry, #surprised)
- crawl tweets during 8 months
- filter very short&misspelled tweets
- use PMI
- investigate influence of negation
  - great [highly positive] $\rightarrow$ not great [mildly negative]
  - terrible [strong negative] $\rightarrow$ not terrible [mildly negative]

Affirmative Context and Negated Context Lexicons

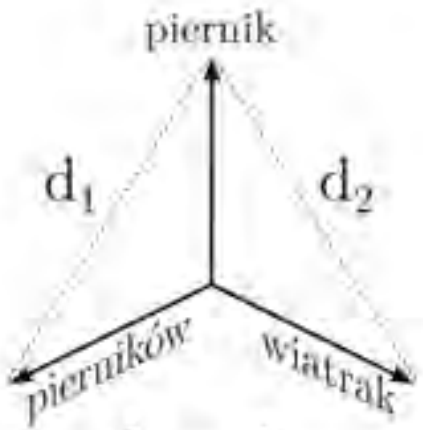
Term	Sentiment140 Lexicons		
	Base	AffLex	NegLex
Positive terms			
great	1.177	1.273	-0.367
beautiful	1.049	1.112	0.217
nice	0.974	1.149	-0.912
good	0.825	1.167	-1.414
honest	0.391	0.431	-0.123
Negative terms			
terrible	-1.766	-1.850	-0.890
shame	-1.457	-1.548	-0.722
bad	-1.297	-1.674	0.021
ugly	-0.899	-0.964	-0.772
negative	-0.090	-0.261	0.389

Affirmative Context and Negated Context Lexicons

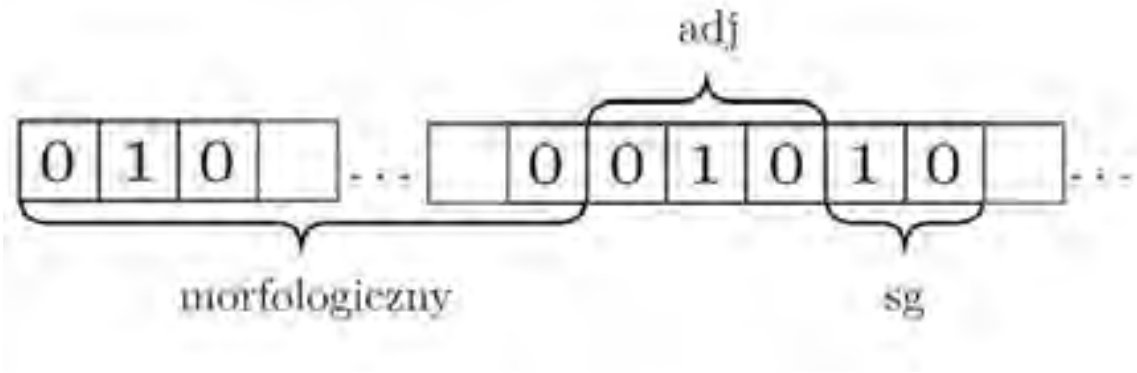


One-hot representation

w	ind(w)	$\boldsymbol{V}(w)$
abakus	1	$[1, 0 \ldots 0]$
leśnikom	820543	$[0 \ldots 1 \ldots 0]$
żyźniejszymi	3624472	$[0 \ldots 0, 1]$



Many-hot representation



Towards dense word representation

The intuition : similar words appear in similar contexts

The cat purrs
This cat hunts mice

The kitty purrs
This kitty hunts mice

The tiger purrs
This tiger hunts mice (?)

Towards dense word representation

The intuition : similar words appear in similar contexts

The cat purrs
This cat hunts mice

The kitty purrs
This kitty hunts mice

The tiger purrs
This tiger hunts mice (?)

Towards dense word representation

The intuition : similar words appear in similar contexts

The cat purrs
This cat hunts mice

The kitty purrs
This kitty hunts mice

The tiger purrs
This tiger hunts mice (?)

Brown Clustering

$$P(corpus|C) = \prod_{i=1}^{n} e\left(w_i|C(w_i)\right) t\left(C(w_i)|C(w_{i-1})\right)$$

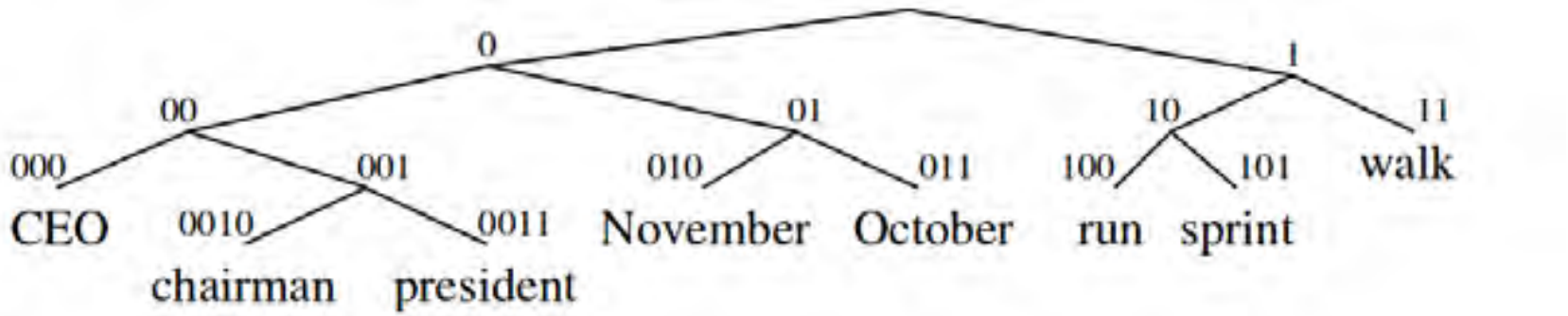
1. Take the top k most frequent words, put each into its own cluster
2. For the rest of words:
   - Create a new cluster for the i^{th} most frequent word
   - Choose two clusters to be merged: pick the merge that gives a maximum value for quality

Brown Clustering

$$P(corpus|C) = \prod_{i=1}^{n} e\left(w_i|C(w_i)\right) t\left(C(w_i)|C(w_{i-1})\right)$$

1. Take the top k most frequent words, put each into its own cluster
2. For the rest of words:
   - Create a new cluster for the i^{th} most frequent word
   - Choose two clusters to be merged: pick the merge that gives a maximum value for quality

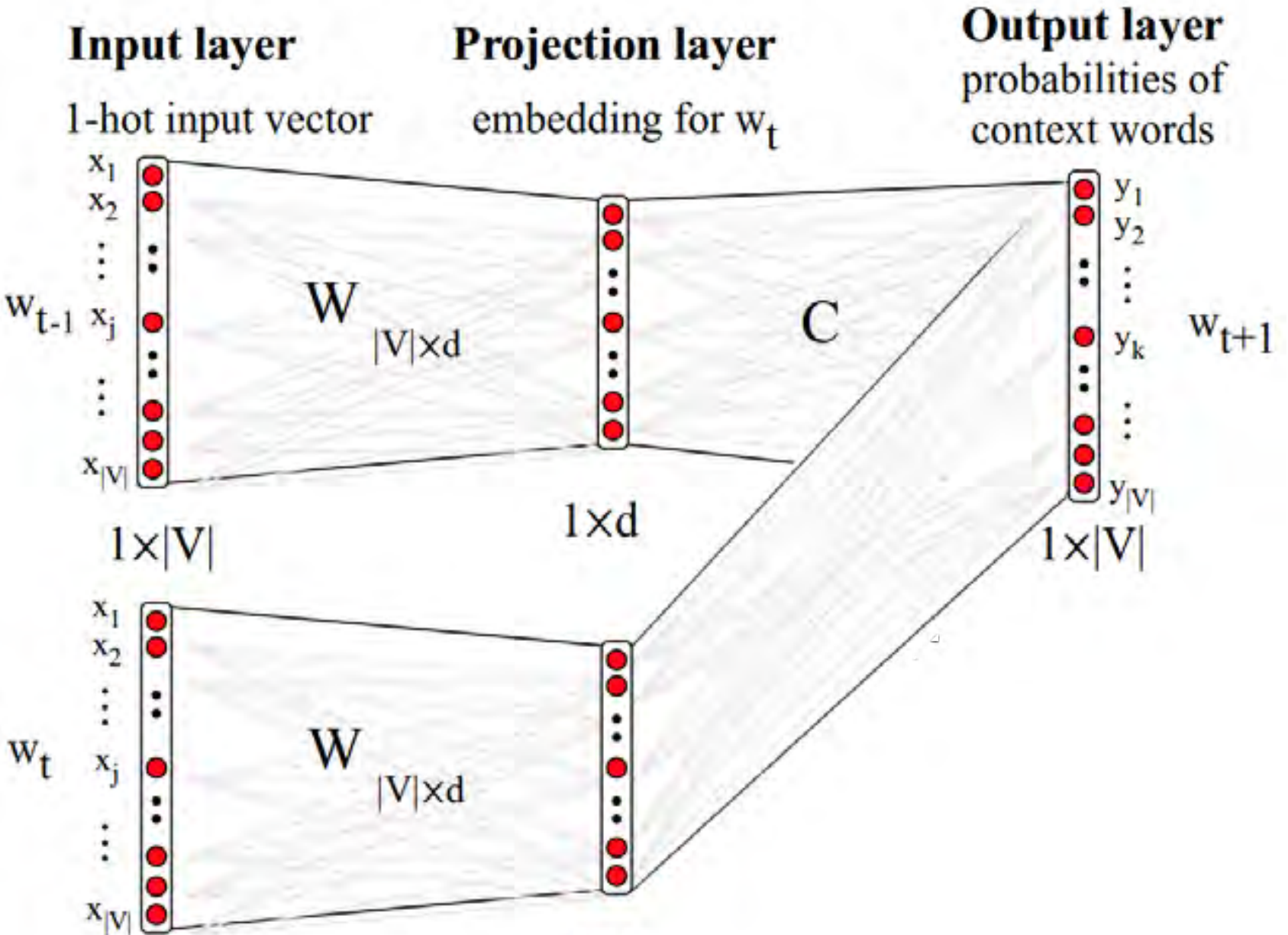
Brown Clustering



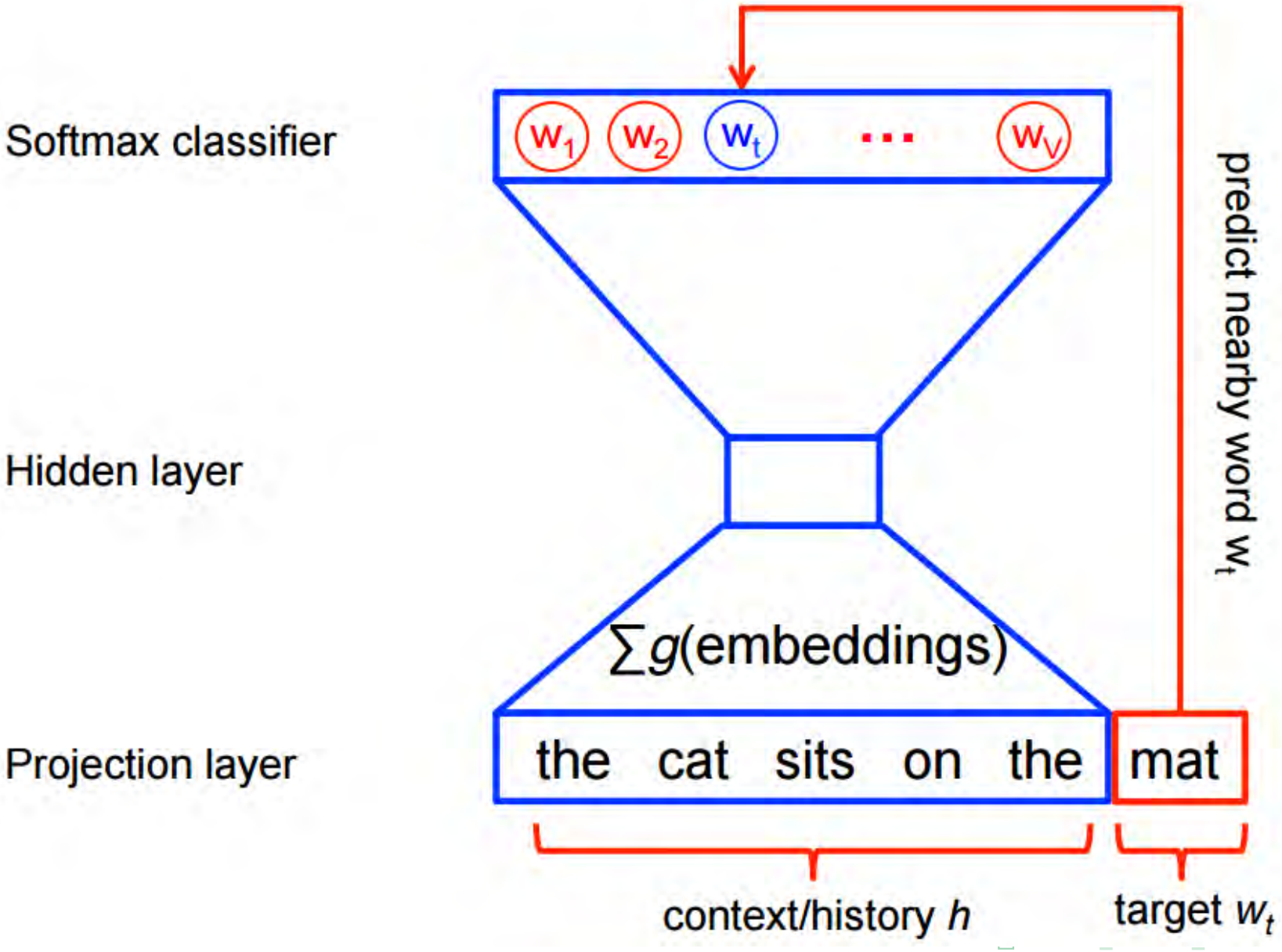
Word embeddings

- Word2Vec
- GloVe

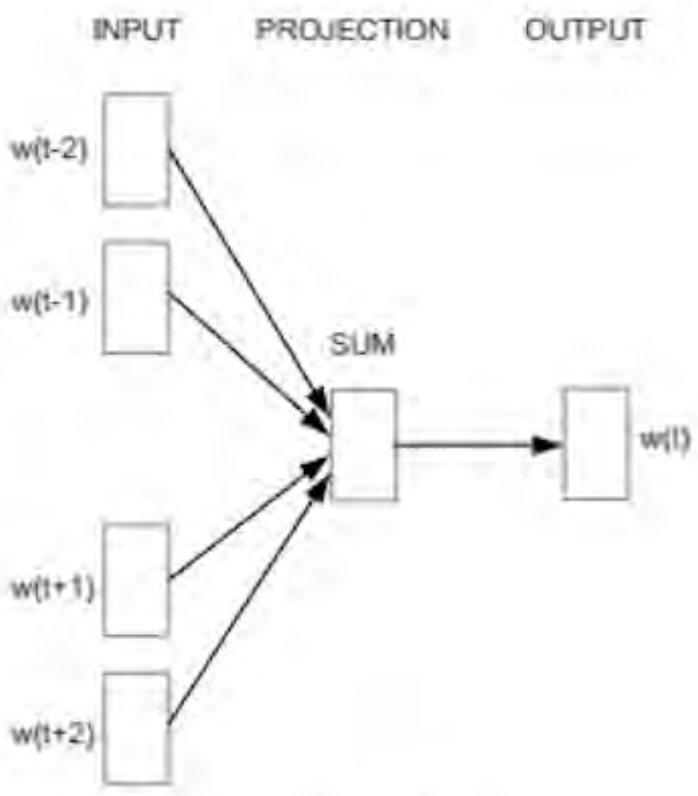
Word2Vec


Input layer
1-hot input vector
Projection layer
embedding for w_t
Output layer
probabilities of context words
w_{t-1}
w_t
w_{t+1}
W
|V|×d
C
1×|V|
1×d
1×|V|

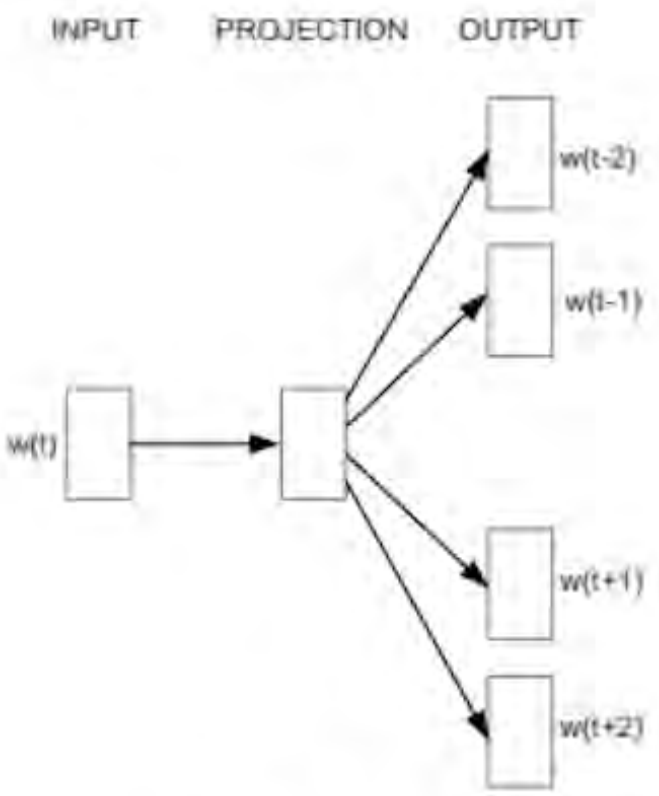
Word2Vec


Softmax classifier
Hidden layer
Projection layer
w_1 w_2 w_t ... w_V
predict nearby word w_t
$\sum g$(embeddings)
the cat sits on the
mat
context/history h
target w_t

Word2Vec



CBOW



Skip-gram

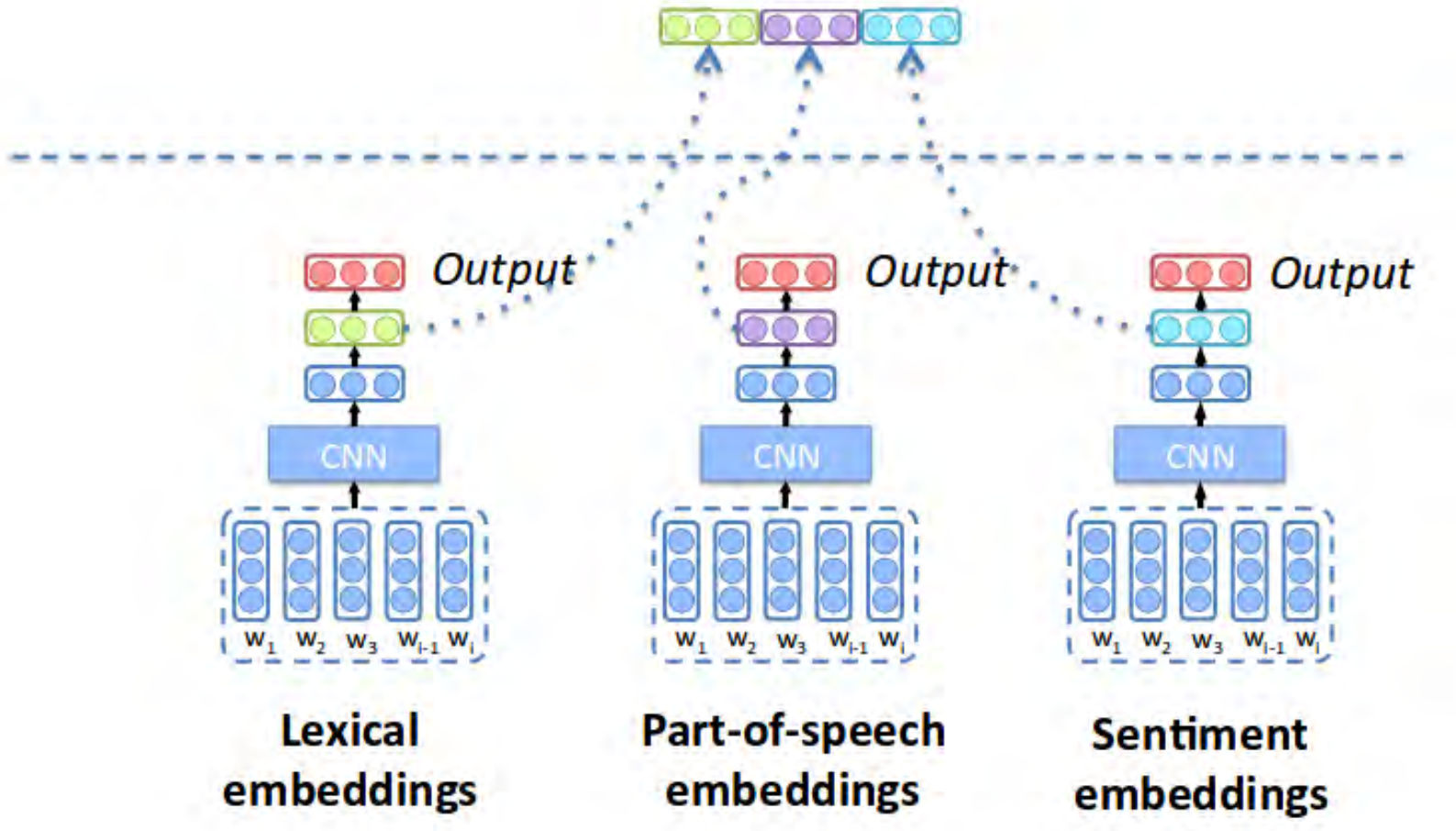
Properties of Word Embeddings

- NN search
  - Japan - Korea, China
  - tea - coffee, lemon, sugar
- semantic analogy
  - puppy - dog $\approx$ kitten - cat
- syntactic analogy
  - taller - tall $\approx$ smaller - small
- "words arithmetic"
  - king - man + woman = ?
  - Paris - France + Germany = ?
  - Tadeusza - Tadeusz + Marek = ?
  - Shakespeare - English + Polish = ?
  - 0.5 (first + fifth) = ?

... embeddings

- Word/Lexical embeddings
- Part-of-speech embeddings
- Document embeddings
- Sentiment embeddings

Polarity embeddings



Classic methods

- Information Gain
- χ^2
- F statistics

Intelligent Feature Selection

- assign initial weight for $a_x = (a_{x1}, a_{x2}, ...)$

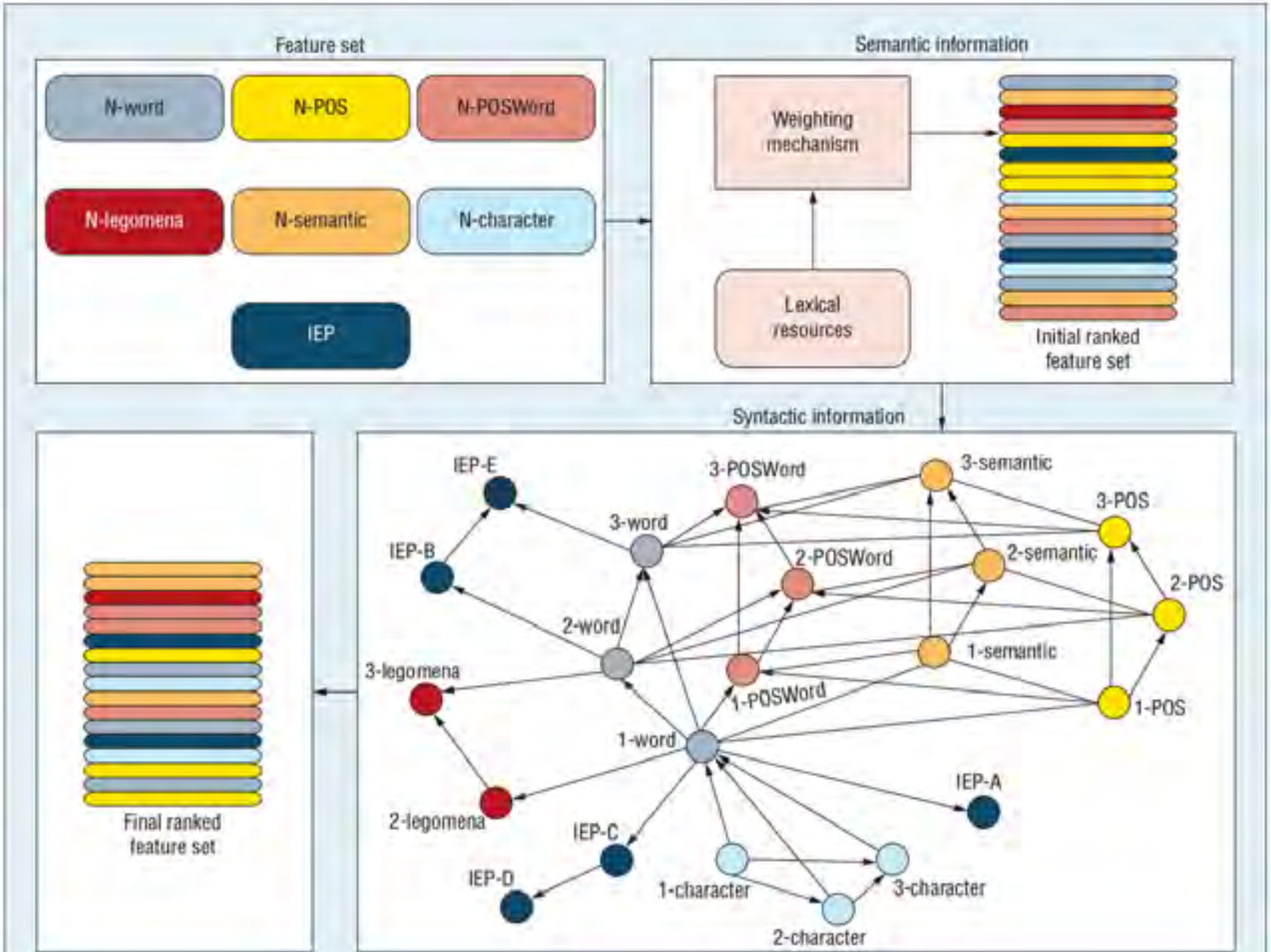
$$w(a_x) = wt(a_x) + ws(a_x)$$

$$wt(a_x) = \max_{v,w} P(a_x|v) \log \frac{P(a_x|v)}{P(a_x|w)}$$

$$ws(a_x) = \frac{1}{d} \sum_{i=1}^{d} \left(\frac{1}{k} \sum_{j=1}^{k} s_{positive}(a_{xi,j}) + s_{negative}(a_{xi,j}) \right)$$

- explore subsumption and parallel relations

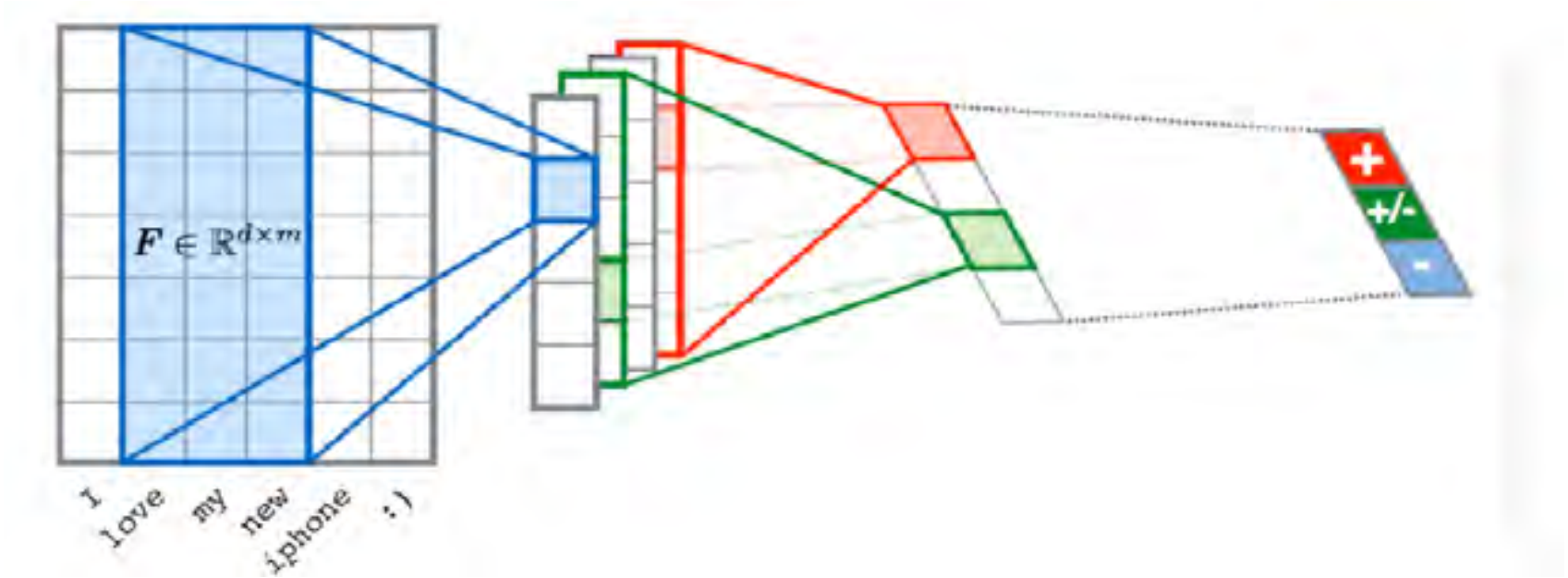
Intelligent Feature Selection



Popular algorithms

- Support Vector Machines
- Random Forests
- Deep Learning...

Convolutional Neural Networks



SemEval 2016 Task 4

- International Workshop on Semantic Evaluation 2016 collocated with the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)
- 10th edition
- 14 different task
- Task 4: Sentiment Analysis in Twitter
  - 4th edition
  - the highest number of participant
  - 43 teams, 25 countries

Our system

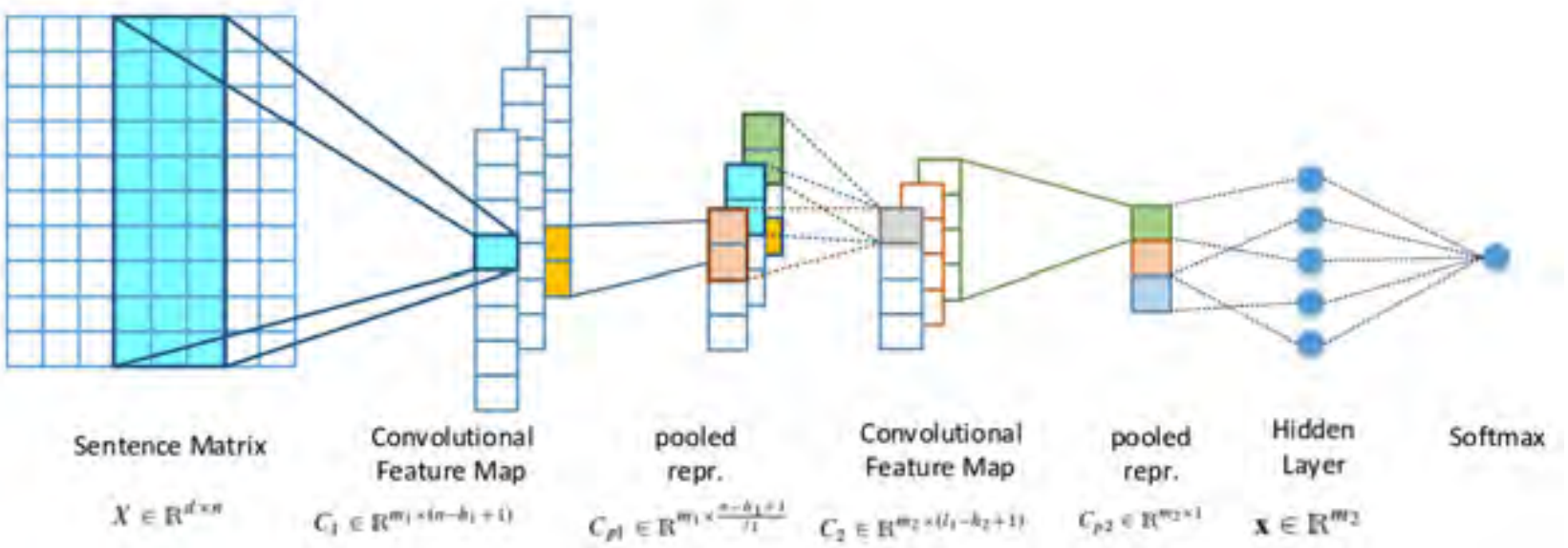
- n-grams, k-grams, negation n-grams, POS-grams
- lexicons: the NRC emotion lexicon , Hu and Liu Opinion lexicon , the Multi-perspective Question Answering corpus , and SentiWordNet
- Hashtag Lexicon
- Brown Clustering
- Gradient Boosting Trees with weights
- SVM, RF added for robustness

Results 4A: Message polarity classification

			F_1^{PN}
1	ETH Zürich	Switzerland	0.633
2	Aix-Marseille University	France	0.630
3	University of Melbourne	Australia	0.617
4	Universidade de Lisboa	Portgual	0.610
5	Athens University of Economics and Business	Greece	0.605
6	Aix-Marseille University	France	0.598
7	Nanyang Technological University	Singapore	0.596
...			
14	Poznan University of Technology	Poland	0.574
...			

Winning algorithm

- 90M tweets (approx. 7K)
- testing set from previous edition
- 2×CNN + RF:
  - Word2Vec ($d = 52$, skip-gram 5, 200M tweets)
  - GloVe ($d = 50$, 90M tweets)



Our system

Algorithm 1 Roughly Balanced Bagging

Input: $D = D_{min} \cup D_{maj}$: original training set of examples of size N, k: number of bootstrap samples, LA: learning algorithm;

Output: C^* bagging ensemble with k component classifiers

1: **for** $i = 1 \rightarrow k$ **do**
2: $N_i^{min} \leftarrow |N_{min}|$
3: $N_i^{maj} \leftarrow$ following negative binomial distribution with $n = N_i^{min}$ and $p = q = 0,5$
4: $S_i^{min} \leftarrow N_i^{min}$-element sample drawn with replacement from D_{min}
5: $S_i^{maj} \leftarrow N_i^{maj}$-element sample drawn with replacement from D_{maj}
6: $C_i \leftarrow LA(S_i^{min} \cup S_i^{maj})$
7: **end for**

$$C^*(x) = \arg\max_y \sum_{i=1}^{k} p_{C_i}(y|x)$$

Results 4B: classification according to a two-point scale

			$recall_{macro}$
1	National Technical University of Athens, University of Athens et al.	Greece	0,797
2	Universidade da Coruna & Universidade de Vigo	Spain	0,791
3	Amazon.in	India	0,784
4	East China Normal University	China	0,768
5	INSIGHT Research Centre, National University of Ireland	Ireland	0,767
6	Poznan University of Technology	Poland	0,763
7	University of Melbourne	Australia	0,758
...			
14	ETH Zürich	Switzerland	0,648
...			

Our system

- "Simple Ordinal" ensemble
- SVM + GBT

Results 4C: classification according to a five-point scale

			MAE_{macro}
1	University of Grenoble-Alpe	France	0,719
2	East China Normal University	China	0,806
3	Poznan University of Technology	Poland	0,860
4	Universidade da Coruna & Universidade de Vigo	Spain	0,864
5	Saints Cyril and Methodius University, Skopje	Macedonia	0,869
6	INSIGHT Research Centre, National University of Ireland	Ireland	1,006
7	Istituto di Scienza e Tecnologie dell'Informazione	Italy	1,074

Feature name	Rel. impor. [%]
NRC Hashtag Lexicon: mean	0.79
Brown cluster: 01110110	0.73
SentiWordNet: sum of negative	0.63
5 k-gram: “d &am”	0.55
Brown cluster: 1110011001111	0.49
NRC Hashtag Lexicon: max	0.48
Opinion Lexicon: negative	0.47
Brown cluster: 111101011101	0.42
3 k-gram “ok ”	0.41
4 k-gram “ nor”	0.40
Brown cluster: 0100100	0.38
3 k-gram “ NY”	0.35
2 n-gram: not against	0.35
Brown cluster: 111101111100100	0.34
5 k-gram “ Anth”	0.34

Tablica: Relative feature importances (%) of top 15 features.

Feature group	Rel. impor. [%]
5 character-gram	26.03
4 character-gram	21.75
3 character-gram	21.74
Brown clusters	6.92
Negated 1-gram	6.62
1-gram + POS	4.24
Negated + 2-gram	3.48
1-gram	2.69
2-gram	1.87
NRC Hashtag Lexicon	1.49
SentiWordNet	1.00
NRC Lexicon	0.93
Opinion Lexicon	0.62
3-gram	0.34
MPQA corpus	0.25
4-gram	0.03

Tablica: Relative feature importances (%) for features groups.

Open challenges

- blending theories of emotions with the practical engineering
- multimodal data
- ordinal classification
- quantification

Quantification

- perfect classifier = perfect quantifier

	T	F
T	80	0
F	0	20

- better classifier ? better quantifier

	T	F
T	70	10
F	10	10

	T	F
T	75	5
F	0	20

Thank you!