University of Stuttgart
Institute for
Natural Language Processing

Emotion Analysis

Machine Learning-based
Classification
Nov 29, 2022

Roman Klinger

@roman_klinger romanklinger
http://www.romanklinger.de/

Outline

1. Recap
2. Introduction
3. ML Methods
   - Feature-based Machine Learning
   - Neural Network-based Approaches
4. Weak and Distant Labeling
   - Obtaining Automatically Annotated Corpora
   - Transfer Learning
5. Multi-task learning
6. Zero-Shot Learning

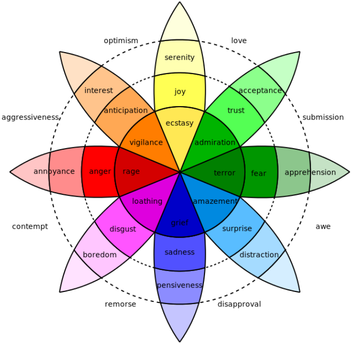
Outline

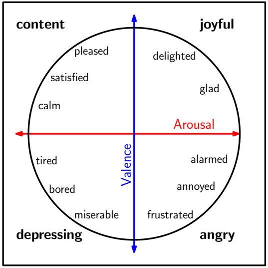
1. Recap
2. Introduction
3. ML Methods
   - Feature-based Machine Learning
   - Neural Network-based Approaches
4. Weak and Distant Labeling
   - Obtaining Automatically Annotated Corpora
   - Transfer Learning
5. Multi-task learning
6. Zero-Shot Learning

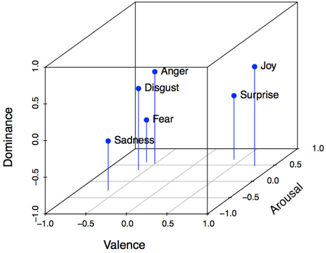
Previous Lectures

- What are emotions? What is the relation to affect?
  How can emotions be organized in psychological models?
- Which annotated corpora exist for emotions?
  How can they be created?
- Which dictionaries exist for emotions?
  How can they be created?

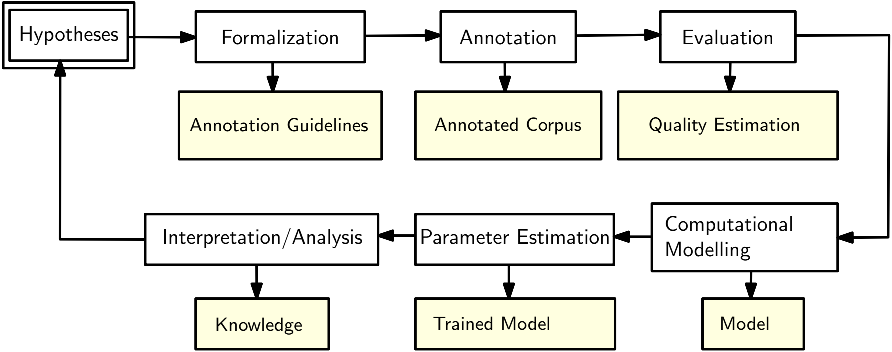
Emotion Models







The Need for Corpora



Corpora

Dataset	Type	Annotation	Size	Source	Avail.
AffectiveText		+ {valence}	1,250	Strapparava (2007)	D-U
Blogs		+ {mixed, noemo}	5,025	Aman (2007)	R
CrowdFlower		+ {fun, love, …}	40,000	Crowdflower (2016)	D-U
DailyDialogs			13,118	Li et al. (2017)	D-RO
Electoral-Tweets			4,058	Mohammad (2015)	D-RO
EmoBank			10,548	Buechel (2017)	CC-by4
EmoInt		- {disgust, surprise}	7,097	Mohammad (2017)	D-RO
Emotion-Stimulus		+ {shame}	2,414	Ghazi et al. (2015)	D-U
fb-valence-arousal			2,895	Preoţiuc (2016)	D-U
Grounded-Emotions		☺☹	2,585	Liu et al. (2017)	D-U
ISEAR		+ {shame, guilt}	7,665	Scherer (1997)	GPLv3
Tales			15,302	Alm et al. (2005)	GPLv3
SSEC			4,868	Schuff et al. (2017)	D-RO
TEC		+ {±surprise}	21,051	Mohammad (2012)	D-RO

Bostan/Klinger, COLING 2018

Dictionaries

- LIWC:
  4500 Words with 80 Psychological Categories
- WordNet Affect:
  ≈ 5000 words, 2000 manually annotated
- NRC Emotion Dictionary:
  ≈ 8000 words, labeled with crowdsourcing for emotions
- NRC VAD:
  ≈ 20000 words, labeled with crowdsourcing via BWS
- DepecheMood:
  ≈ 37000 words, weakly annotated
- ANEW:
  ≈ 1600 words, expert annotated for VAD

Take Away

- Motivation for emotion classification
- Approaches for emotion classification
  - Dictionaries, Features, Neural
  - Weak/Distant Supervision, Transfer Learning
  - Multitask learning
  - Zero-shot learning

Outline

1. Recap
2. Introduction
3. ML Methods
   - Feature-based Machine Learning
   - Neural Network-based Approaches
4. Weak and Distant Labeling
   - Obtaining Automatically Annotated Corpora
   - Transfer Learning
5. Multi-task learning
6. Zero-Shot Learning

Task Definition

Emotion classification is the task to assign one or multiple emotions from a predefined emotion inventar to a textual unit, e.g., a document, paragraph, sentence.

Dictionary-based Methods

First attempt: Use a dictionary D_e of entries t with emotion scores $s_e(t)$ for emotion e:

$$\text{score}(\text{text}, e) = \frac{1}{|\text{text}|} \sum_{w \in \text{text}} s_e(w)$$

- Issues? Number of words in dictionary associated with emotion might differ. Normalize:

$$\text{score}(\text{text}, e) = \frac{1}{|D_e|} \frac{1}{|\text{text}|} \sum_{w \in \text{text}} s_e(w)$$

- Decision for an emotion:

$$\text{emotion}(\text{text}) = \arg\max_{e \in \text{Emotions}} \text{score}(\text{text}, e)$$

Limitations of Dictionaries?

Limitations of Dictionaries?

- Dictionaries alone might not capture intensifiers, modifiers, negations
- Irony or sarcasm, figurative language
- Implicit formulations
- References to events
- Coverage and precision might be limited
- Compositionality is not captured
- Not all emotion expressions might be captured equally well
- ...

Machine learning based

Statistical methods

- Advantages:
  - Model adapts well if training data available
  - Might capture aspects that are difficult to encode in rules
  - Might capture aspects that we do not know
- Disadvantages:
  - Corpus is required
  - Might not adapt well to domains outside of training data
  - Not necessarily transparent

Outline

1. Recap
2. Introduction
3. ML Methods
   - Feature-based Machine Learning
   - Neural Network-based Approaches
4. Weak and Distant Labeling
   - Obtaining Automatically Annotated Corpora
   - Transfer Learning
5. Multi-task learning
6. Zero-Shot Learning

Which features could be helpful?

Features in a ML setting?

- I am happy
- I am not happy
- He is in Disneyland
- She has a date
- That sucks
- Sure. That's reaaaallllly cooool!!1
- She got a new bicycle
- She needed to buy a car
- His son ran on the street!
- I did not pass the exam.

Alm 2005: Tales

- Learn function to map sentence to emotion
  - Neutral vs. Emotional
  - Neutral vs. Positive vs. Negative
- Features: First sentence, conjunctions, quoted, story type, special punctuation, complete upper-case words, sentence length, range of story progress, number of JJ, N, V, RB, verb count, positive/negative count, word net emotion words, interjections, bag of words
- Classifier: SNoW (a rule learning method)
- Result:
  - N/E: 70% F_1 for N, 47% for E
  - N/P/NE: 69% F_1 for N, 32% F_1 NE, 13% F_1 P

Strapparava 2007: Affective Text Headlines

- Data set used for shared task
- Three teams participated
- Mostly no machine learning but rules with dictionaries and similarity measures of words to other resources
- Results between 15 and 30 % F_1

Aman 2007

- Features based on words on dictionaries (General Inquirer and WordNet Affect)
- Paper leaves some ambiguities open how these dictionaries were used as features
- Classifiers: Naive Bayes and Support Vector Machine

Features	**Naïve Bayes**	**SVM**
GI	71.45%	71.33%
WN-Affect	70.16%	70.58%
GI+WN-Affect	71.7%	73.89%
ALL	**72.08%**	**73.89%**

Schuff, 2017, Stance Sentiment Emotion Corpus (SSEC)

MaxEnt, Linear SVM

- Bag-of-Words

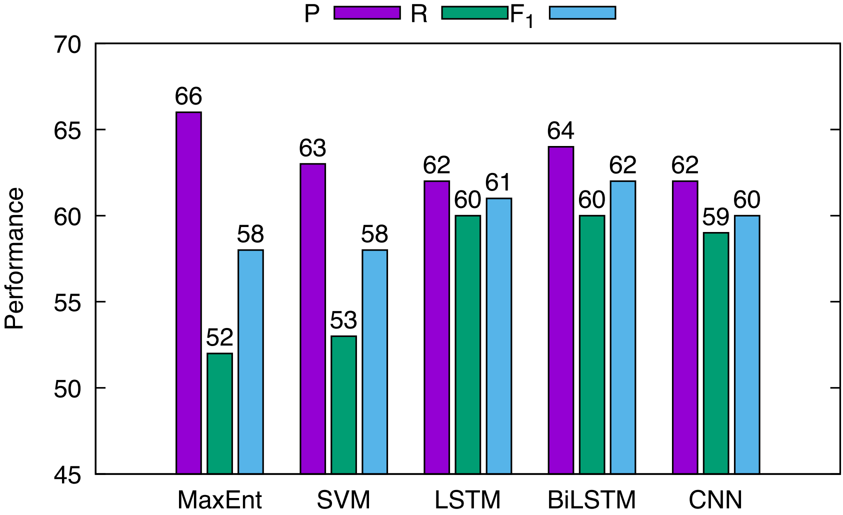
LSTM, BiLSTM

- 300 dimensional embedding
- 175 dimensional LSTM layer, 0.5 dropout rate
- 50 dimensional dense layer
- 8 output neurons

CNN

- Convolution of window size 2,3,4
- Pooling of length 2

Schuff, 2017, Stance Sentiment Emotion Corpus (SSEC)



Outline

Neural Network Architectures

Which architectures make sense?

Shared Tasks

- Affective Text (Headlines), 2007 (SemEval)
- Emotion Intensity, 2017 (WASSA), 2018 (SemEval)
  (not discussed here, see session on intensity next year)
- Emotion Classification (E-c) 2018 (SemEval)
- Implicit Emotions, 2018 (WASSA)

Emotion Classification E-c SemEval, Setting

Task Definition

Emotion Classification (E-c): Given a tweet, classify it as ‘neutral or no emotion’ or as one, or more, of eleven given emotions that best represent the mental state of the tweeter

- Annotation via crowdsourcing
- Aggregation:
  Accept emotion label with at least 2/7 annotations

Emotion Classification E-c SemEval

	#Teams				
ML algorithm	EI-reg	EI-oc	V-reg	V-oc	E-c
AdaBoost	1	1	3	1	0
Bi-LSTM	10	8	10	6	6
CNN	10	8	7	6	3
Gradient Boosting	8	3	5	4	1
Linear Regression	11	2	7	2	1
Logistic Regression	9	7	8	6	6
LSTM	13	9	10	5	4
Random Forest	8	7	5	6	6
RNN	0	0	0	0	1
SVM or SVR	15	9	8	6	6
Other	14	16	13	12	7

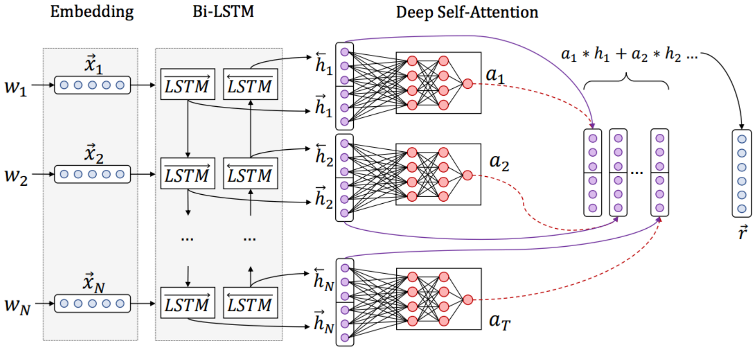
Figure 2: Machine learning algorithms used by teams.

Emotion Classification E-c SemEval

Rank	Team Name	acc.	micro F1	macro F1
English				
1	NTUA-SLP	58.8	70.1	52.8
2	TCS Research	58.2	69.3	53.0
3	PlusEmo2Vec	57.6	69.2	49.7
17	Median Team	47.1	59.9	46.4
21	SVM-Unigrams	44.2	57.0	44.3
28	Random Baseline	18.5	30.7	28.5
Arabic				
1	EMA	48.9	61.8	46.1
2	PARTNA	48.4	60.8	47.5
3	Tw-StAR	46.5	59.7	44.6
6	SVM-Unigrams	38.0	51.6	38.4
7	Median Team	25.4	37.9	25.0
9	Random Baseline	17.7	29.4	27.5
Spanish				
1	MILAB_SNU	46.9	55.8	40.7
2	ELiRF-UPV	45.8	53.5	44.0
3	Tw-StAR	43.8	52.0	39.2
4	SVM-Unigrams	39.3	47.8	38.2
7	Median Team	16.7	27.5	18.7
8	Random Baseline	13.4	22.8	21.3

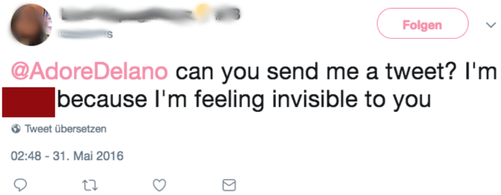
Table 11: Task 5 emotion classification (E-c): Results.

SemEval E-c SemEval Winner



https://www.aclweb.org/anthology/S18-1037

Implicit Emotions Shared Task: Data and Task

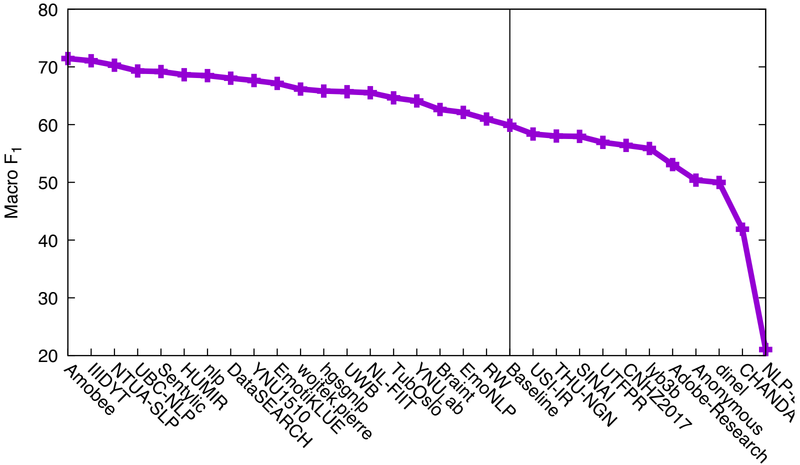


- Input:
  Tweet with emotion synonym replaced by unique string
- Output:
  Emotion for which the removed word is a synonym

Example

```
sadness [USERNAME] can you send me a tweet? I'm
[#TRIGGERWORD#] because I'm feeling invisible to you
```

Implicit Emotions Shared Task: Results



Implicit Emotions Shared Task: Tools

- Deep learning:
  - Keras, Tensorflow
  - PyTorch of medium popularity
  - Theano only once
- Data processing, general ML:
  - NLTK, Pandas, ScikitLearn
  - Weka and SpaCy of lower popularity
- Embeddings/Similarity measures:
  - GloVe, GenSim, FastText
  - ElMo less popular

Implicit Emotions Shared Task: Methods

- Nearly everybody used embeddings
- Nearly everybody used recurrent neural networks (LSTM/GRU/RNN)
- Most top teams used ensembles (8/9)
- CNNs distributed ≈ equally across ranks
- Attention mechanisms 5/9 top, not by lower ranked teams
- Language models used by 3/4 top teams
- Winner: `https://www.aclweb.org/anthology/W18-6207/`
- More information: `https://www.aclweb.org/anthology/W18-6206/`

Outline

Weak and Distant Labeling

Weak/Distant Labeling:

- Use an external authority for labeling instances
- Example:
  - Use data base with entities for NER annotation
  - Use data base with relations for relation annotation
  - Use dictionary of emotion words to label instances
  - Commonly used for EA on social media:

Self Labeling:

- Predict an element from the text from the rest
- Examples:
  - Predict an emoji, emoticon
  - Predict hashtag
  - Predict a word (as in Implicit Emotions Shared Task)

Self-Labeling

Approach:

- Manually associate
  - hashtags with emotions
  - emojis with emotions
- Assume that occurrence of hashtag/emoji marks emotion
- Predict "self-labeled emotion" from text after removing hashtag/emoji
- Apply to other texts

Advantage:

- Easy to obtain huge data sets

Disadvantage:

- Concept of emotion $\neq$ emotion hashtags/emojis
- Example: 10.1109/SocialCom-PASSAT.2012.119

Outline

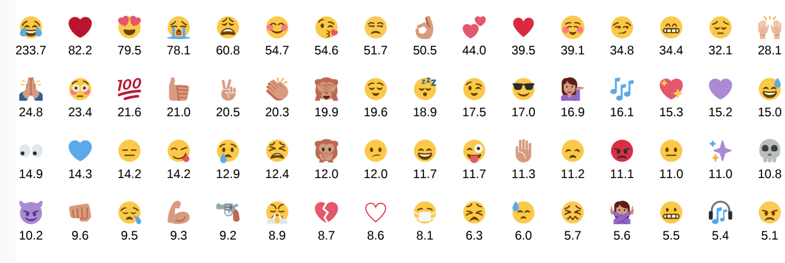
Transfer Learning

- Idea:
  - Pretrain model on related task (where data is easy to get)
  - Fine-tune model on actual task
- Challenge: Catastrophic Forgetting
- Related methods:
  - Domain adaptation

General Approaches to Transfer Learning

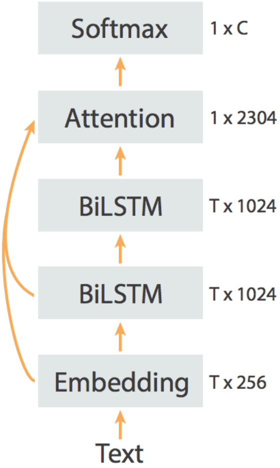
- ULMFit
  - Pretrain LSTM-based language model
  - Fine-tune to specific task
  - Tackles catastrophic forgetting by gradual unfreezing
- BERT
  - Transformer architecture
  - Pretrain joint sentence and contextualized embeddings
  - Fine-tune top layers on specific task
- Embeddings, like Word2Vec
  - Predict a word from context
  - Predict context from word
  - Use representations to start a classifier with, fine-tune embeddings for task

Transfer Learning: DeepMoji

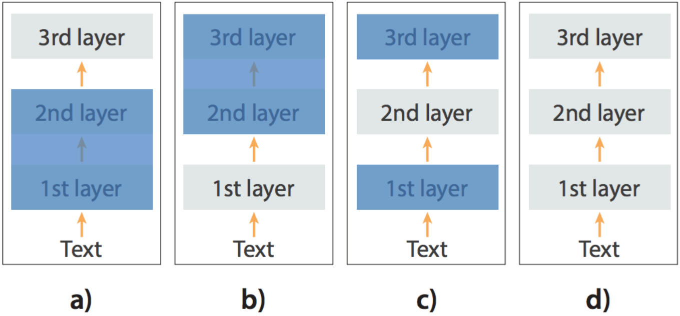


Transfer Learning: DeepMoji

- Develops a deep learning method for emotion classification (amongst other tasks)
- Pretrain model on huge data set to predict the occurrence of an emoji
- Fine-tune: Keep subset of parameters fixed while learning on actual data set.



Transfer Learning: DeepMoji



- Blue: frozen
- a) tune any new layers
- b) then tune 1st layer
- c) then tune next layer, until all have been tuned
- d) tune all together

DeepMoji Demo and Reference

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, Sune Lehmann: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. EMNLP 2017.

- Demo: `https://deepmoji.mit.edu/`
- Paper: `https://aclanthology.org/D17-1169/`

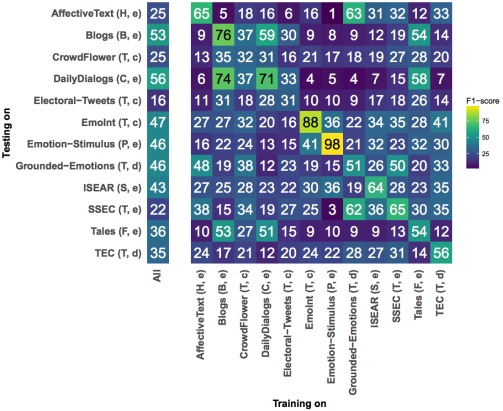
Final Remark on Results

- Results differ a lot between data sets
- Data sets are pretty incomparable
- Do not assume a high number is a good result or a low number is a bad result without understanding the data set.

Final Remark on Results

Cross-corpus experiment

- Split corpora in train/val
- Train BOW-MaxEnt-L2 on all train parts, apply on all val parts
- Join all train parts, apply on each val part



(https://www.aclweb.org/anthology/C18-1179/)

Outline

Obervations

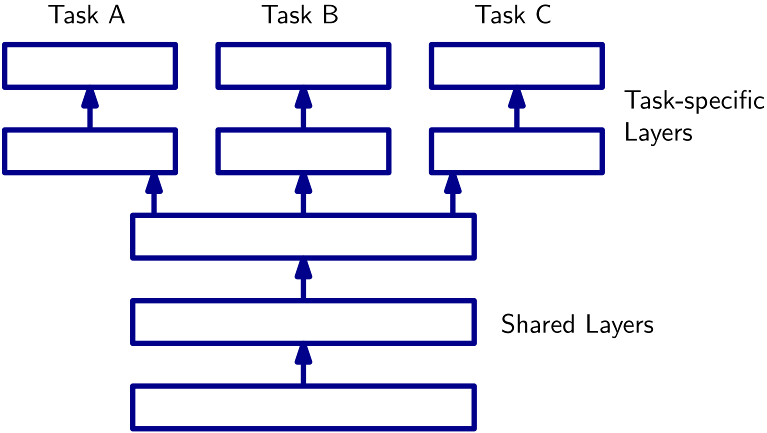
Task specific developments

- Emotion dictionary features
- Handle intensifiers/negations
- Adapt input representations (e.g., retrofitting)
- Pretraining of particularly relevant proxy task

General developments

- Neural networks work best
- Methods that work well across different classification tasks work well for emotion analysis

Overview of Multitask learning



Tasks in Multitask Learning and Emotions

- Akhtar et al, NAACL 2019: Multi-task Learning for Multi-modal Emotion Recognition and Sentiment Analysis
  https://www.aclweb.org/anthology/N19-1034.pdf
- Chauhan et al, ACL 2020: Sentiment and Emotion help Sarcasm? A Multi-task Learning Framework for Multi-Modal Sarcasm, Sentiment and Emotion Analysis
  https://www.aclweb.org/anthology/2020.acl-main.401.pdf
- Dankers et al, EMNLP 2019: Modelling the interplay of metaphor and emotion through multitask learning
  https://www.aclweb.org/anthology/D19-1227.pdf

Tasks in Multitask Learning and Emotions

- Tafreshi et al, CoNLL 2018: Emotion Detection and Classification in a Multigenre Corpus with Joint Multi-Task Deep Learning
  https://www.aclweb.org/anthology/C18-1246.pdf
- Rajamanickam et al, ACL 2020: Joint Modelling of Emotion and Abusive Language Detection
  https://www.aclweb.org/anthology/2020.acl-main.394.pdf
- Saha et al, ACL 2020: Towards Emotion-aided Multi-modal Dialogue Act Classification
  https://www.aclweb.org/anthology/2020.acl-main.402.pdf
- Casel et al, KONVENS 2021: Emotion Recognition under Consideration of the Emotion Component Process Model.
  https://aclanthology.org/2021.konvens-1.5/

Summary

- Feature-based emotion analysis research came up with rich sets of task-specific properties
- Deep learning, transfer learning outperforms such approaches mostly, but is sometimes also combined.
- Current research is a lot about finding beneficial proxy tasks and to adapt input representations

Outline

Recent developments; Zero-Shot Learning for Emotion Classification

- Sometimes, dictionary-based emotion classification is called unsupervised
  - That is obviously wrong, if supervision went into the dictionary-creation process.
  - Still, this term highlights a desideratum: assign labels without training data.
- Similar: Given development data with emotion labels, develop a model, that can predict unseen labels.

  ⇒ Zero-Shot Learning

Why should Zero-Shot Learning be possible?

Training Data with labels: Deer, Fish, Rabbit

Deer

Fish

Rabbit

Test Data with unseen labels: Moose, Whale

Moose

Whale

- How do we make these assignments?
- We decide on properties of the instances to classify.
- We compare the extracted properties to those of the classes.
- We need some kind of feature vector/embedding of each instance and label.

Feature-based Generalization

Features: (eats-grass, has-lungs, lives-in-water, is-cute)

Labels during training: Deer, Fish, Rabbit

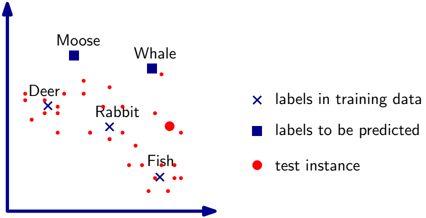
Deer	Fish	Rabbit
(1,1,0,0)	(0,0,1,0)	(1,1,0,1)

Unseen labels: Moose, Whale

Moose	Whale
(1,1,0,1)	(0,1,1,0)

⇒ If we had a model which predicted these feature vectors, we could just select the most appropriate label based on a neighest-neighbor approach

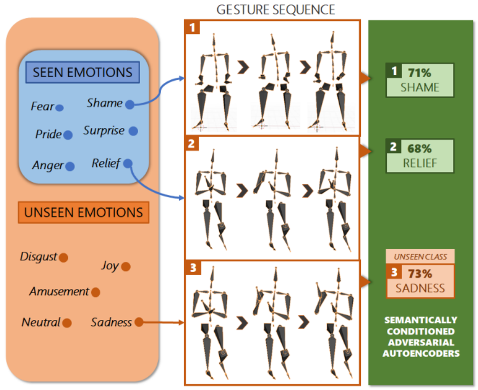
ZSL as Embedding Prediction



- Vectors based on concept features
- Get feature vector for a new instance

- In ZSL, we would assign "whale".
- In Generalized ZSL, we assign "fish".
- Hubness problem: It's more likely to predict vectors that have been seen at model development time.
- Emotion analysis: compare text-embeddings (instances) to emotion-name embeddings (as above)

Related: ZSL for Emotion Classification from Gestures



`https://aaai-2022.virtualchair.net/poster_aaai10434`

ZSL for Emotion Classification

- Learn function $f(\text{text}) \rightarrow \overrightarrow{\text{emotion}}$
  - Where to get emotion vector representation from?
  - Word embeddings of emotion names, appraisals, VAD …
  - Vector needs to correspond to an emotion category
- At inference time: embed text to vector, assign nearest emotion (seen or unseen during training)
- Our experiments:
  works ok for seen emotions, but not for others

⇒ Hubness Problem :-(

(work with Flor Miriam Plaza del Arco)

Another approach to ZSL Emotion Classification

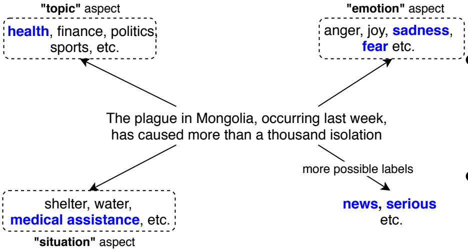
- Recent unpublished work: Chochlakis et al (Oct 2022): Using Emotion Embeddings to Transfer Knowledge between Emotions, Languages, and Annotation Formats. `https://arxiv.org/pdf/2211.00171.pdf`
- Idea: Provide set of emotions at inference time that are to be predicted
- Predefine emotions clusters, neural network predicts cluster embeddings
- Regularize such that similar emotions (according to prior knowledge) are close in parameter space
- I am not sure what happens at inference time,
  but I contacted the authors. I'll update you when I learn from them how it works....

Alternative: Zero-Shot Learning as Entailment

Benchmarking Zero-shot Text Classification:
Datasets, Evaluation and Entailment Approach

Wenpeng Yin, Jamaal Hay, Dan Roth
Cognitive Computation Group
Department of Computer and Information Science, University of Pennsylvania
{wenpeng,jamaalh,danroth}@seas.upenn.edu

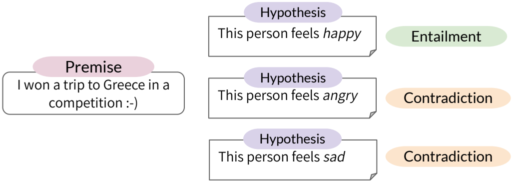
Zero-Shot Learning as Entailment (2)



- Input:
  Two sentences, premise and hypothesis
- Output:
  contradiction, entailment, neutral
- Example online demo:
  `https://huggingface.co/microsoft/deberta-large-mnli`

- How to represent the label as a hypothesis?
- Yin et al. use "This text expresses [?]" and the WordNet concept definition.

Emotion ZSL as Natural Language Inference



- Does it matter which NLI model we use as a backbone?
- How to represent the emotion?
- Should we use multiple emotion representations to increase coverage?

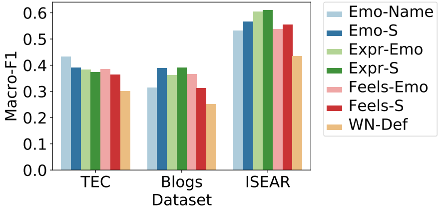
(Arco Del Plaza et al COLING 2022: Natural Language Inference Prompts for Zero-shot Emotion Classification in Text across Corpora)

Emotion Hypotheses

Emo-Name	angry
Expr-Emo	This text expresses anger
Feels-Emo	This person feels anger
WN-Def	This person expresses a strong emotion; a feeling that is oriented toward some real or supposed grievance

Emo-S	Same prefix + anger, annoyance, rage, outr fury, irritation
Expr.-S	
Feels.-S	
EmoLex	all emotion words from an NRC emotion lexicon

The role of the prompt design



(Supervised RoBERTa model: TEC/Blogs: ≈.69, ISEAR: ≈.73)

- TEC: single emotion names work better than with synonyms
- BLOGS: synonyms harm the performance for Feels-Emo/S prompts
- Generally: synonyms help, except for some cases, in which annotaton procedure might be the reason

Use of an emotion lexicon to generate prompt?

	TEC			BLOGS			ISEAR		
Model	P	R	F_1	P	R	F_1	P	R	F_1
d-ensemble	.42	.44	.41	.40	.65	.35	.67	.62	.59
d-emolex	.37	.36	.33	.52	.48	.48	.47	.42	.40
non-zsl	.69	.69	.69	.72	.71	.69	.73	.73	.73

- (d-ensemble is a DeBERTa-based ensemble of all prompts mentioned before)
- Only works for one of our domains: Blogs
- Super-slow (one prompt for every concept in the lexicon)

Observation:

These models are all pretty non-knowledgeable about the concept of emotions.

We will see approaches that explore the structure of emotions and psychological theories more in the next lecture.

Take Away

- Motivation for emotion classification
- Approaches for emotion classification
  - Dictionaries, Features, Neural
  - Weak/Distant Supervision, Transfer Learning
  - Multitask learning
  - Zero-shot learning

University of Stuttgart
Institute for
Natural Language Processing

Emotion Analysis

Machine Learning-based
Classification
Nov 29, 2022

Roman Klinger

@roman_klinger romanklinger
http://www.romanklinger.de/