Sentiment Analysis
State of the Art in Research and Industry

Mark Cieliebak
Zurich University of Applied Sciences
Mark Cieliebak

+ PhD in Theoretical Computer Science
+ IT Consultant in Major Swiss Bank
+ CIO at Netbreeze (bought by Microsoft)
+ >30 Publications

Lecturer

CEO

Conference Chair
Sentiment Analysis

**Goal:** Decide whether a text expresses positive or negative emotion.

"This is a nice conference!"
Insights for Marketing and Sales

Sentiment Analysis can identify trends in Social Media
Characteristics of Sentiment Analysis

Labels:
- Positive
- Negative
- Neutral
- Mixed
- (unknown)

Tasks:
- Single sentence
- Complete document
- Specific aspect/target
- Quantification
Sentiment-Analysis sounds easy

…but it isn't

@francesco_con40 2nd worst QB. DEFINITELY Tony Romo. The man who likes to share the ball with everyone. Including the other team

Tim Tebow may be available! Wow Jerry, what the heck you waiting for! http://t.co/a7z9FBL4

@prodnose is this one of your little jokes like Elvis playing at the Marquee next Tuesday?

#YouCantDateMe if u still sag ur pants super hard...dat shit is played the fuck out!!!
A Remark about Tool Quality

"They all suck...and we suck, too."

CEO of a sentiment analysis company (2013)
Evaluation of Commercial Sentiment Analysis Tools in 2013

7 Text Corpora

- Single statements
- Various media types (tweet, news, reviews, speech transcripts etc.)
- Total: 28'653 texts

9 Commercial APIs

- Stand-alone
- Free for this evaluation
- English text
Quality of Commercial Tools in 2013

Quality of Commercial Tools in 2013

SemEval: International Competition for Sentiment Analysis

Task: Build a system for sentiment analysis (pos, neg, neutral) on tweets in English

<table>
<thead>
<tr>
<th>Year</th>
<th>Winning Team</th>
<th>F1-Score</th>
<th>Winning Technology</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>NRC Canada</td>
<td><strong>69.02</strong></td>
<td>Features + large dictionaries</td>
<td>First run of the competition</td>
</tr>
<tr>
<td>2014</td>
<td>TeamX</td>
<td><strong>72.12</strong></td>
<td>Similar approach as in 2013</td>
<td>First two participants using deep learning</td>
</tr>
<tr>
<td>2015</td>
<td>Webis</td>
<td>64.84</td>
<td>Ensemble of 4 approaches from previous years</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>SwissCheese</td>
<td><strong>63.30</strong></td>
<td>CNN+Distant Supervision</td>
<td>30'000 new tweets Dominance of deep learning among submissions</td>
</tr>
</tbody>
</table>
Did Sentiment Technology Improve?

F1-Score

- Winner of the Respective Year
Did Sentiment Technology Improve?

F1-Score

2013 2014 2015 2016

Winner of the Respective Year

Winner of 2016

Red line: performance of SemEval winner from 2016 (SwissCheese) if only trained on training data for each year

Winner of SemEval 2016 would have won all previous competitions
A Shallow Dive into Technology
SwissCheese: 3-Phase Training with Distant Supervision

Twitter

Raw Tweets (200M) → word2vec

Smiley Tweets (90M) → Distant Supervision

Annotated Tweets (18k) → 2-Layer CNN

Word Embeddings

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>cats</th>
<th>cute</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>0.1</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>cats</td>
<td>0.3</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>cute</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Adapted Word Emb.

<table>
<thead>
<tr>
<th></th>
<th>cat</th>
<th>cats</th>
<th>cute</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td>0.1</td>
<td>0.9</td>
<td>0.3</td>
</tr>
<tr>
<td>cats</td>
<td>0.3</td>
<td>0.2</td>
<td>0.7</td>
</tr>
<tr>
<td>cute</td>
<td>0.2</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Predictive Model

Unknown Tweet
2-Layer Convolutional Neural Network for Sentiment Analysis

Convolution (1) with added bias and activation function

Max Pooling (1)

Convolution (2) with added bias and activation function

Max Pooling (2)

Softmax

\[ S \in \mathbb{R}^{d \times n} \]

\[ C_1 \in \mathbb{R}^{m_1 \times n + h_1 - 1} \]

\[ C_{pool 1} \in \mathbb{R}^{m_1 \times q} \] (transposed visualization)

\[ C_2 \in \mathbb{R}^{m_2 \times q + h_2 - 1} \]

\[ C_{pool 2} \in \mathbb{R}^{m_2 \times 1} \]

\[ P(y = j) \in \mathbb{R}^{K} \]

Intermediary Results

Network Parameters

\[ F_1 \in \mathbb{R}^{m_1 \times d \times h_1} \]

\[ b_1 \in \mathbb{R}^{m_1} \]

\[ F_2 \in \mathbb{R}^{m_2 \times m_1 \times h_2} \]

\[ b_2 \in \mathbb{R}^{m_2} \]

\[ W \in \mathbb{R}^{m_2 \times K} \]
Distant Phase rearranges Word Embeddings

Before the Distant Phase

After the Distant Phase
The More Data, The Better!

Number of tweets in distant phase

Number of annotated tweets
Learn on Tweets, Classify News?

<table>
<thead>
<tr>
<th></th>
<th>train</th>
<th>test</th>
<th>SemEval'13_tweets</th>
<th>MPQ_reviews</th>
<th>DIL_reviews</th>
<th>DAI_tweets</th>
<th>Union of All Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>SemEval'13_tweets</td>
<td></td>
<td></td>
<td>72.4</td>
<td>45.8</td>
<td>53.1</td>
<td>62.2</td>
<td>63.9</td>
</tr>
<tr>
<td>MPQ_reviews</td>
<td></td>
<td></td>
<td>62.2</td>
<td>54.1</td>
<td>40.9</td>
<td>57.8</td>
<td>58.7</td>
</tr>
<tr>
<td>DIL_reviews</td>
<td></td>
<td></td>
<td>57.3</td>
<td>36.8</td>
<td>55.1</td>
<td>48.5</td>
<td>52.9</td>
</tr>
<tr>
<td>DAI_tweets</td>
<td></td>
<td></td>
<td>67.9</td>
<td>37.7</td>
<td>50.4</td>
<td>70.8</td>
<td>60.4</td>
</tr>
<tr>
<td>Union of All Training Data</td>
<td></td>
<td></td>
<td>73.0</td>
<td>50.8</td>
<td>49.9</td>
<td>76.6</td>
<td>66.6</td>
</tr>
</tbody>
</table>

Measured in F1 score

Cross-Domain Performance of SemEval Winner 2016
# Sentiment for other Languages

<table>
<thead>
<tr>
<th>Language</th>
<th>Available Data</th>
<th>Best Know Result (F1 Score)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>German</td>
<td>10,000 Tweets</td>
<td>64.19</td>
<td>Deriu et al., 2016, WSDM (submitted)</td>
</tr>
<tr>
<td>Spanish</td>
<td>68,000 Tweets</td>
<td>71.1 (precision)</td>
<td>Villena-Roman et al., 2013, Procesamiento del Lenguaje Natural</td>
</tr>
<tr>
<td>Italian</td>
<td>7,000 Tweets</td>
<td>65.87</td>
<td>Deriu et al., 2016, WSDM (submitted)</td>
</tr>
<tr>
<td>Dutch</td>
<td>1,100 Tweets (labeled pos/neg)</td>
<td>88.33</td>
<td>Deriu et al., 2016, WSDM (submitted)</td>
</tr>
<tr>
<td>Arabic</td>
<td>1,100 Tweets</td>
<td>73.5</td>
<td>Salab et al., 2015, ANLP</td>
</tr>
</tbody>
</table>
We did it: Theory is over!
Do It Yourself: Sentiment Analysis Tools and APIs

Big Players
• Google Prediction API
• IBM AlchemyAPI
• Microsoft Azure Text Analytics API

NLP Specialists
• RapidMiner
• Repustate
• Semantria
• SentiStrength
• SpinningBytes

Development Toolkits
• Natural Language ToolKit NLTK (Python)
• StanfordNLP (Java)
Understand Customer Reviews

Example: Aspect-based Sentiment Analysis for Hotel Reviews

Use Twitter to predict Heart Disease Mortality

Source: Eichstaedt et al., 2015: Psychological Language on Twitter Predicts County-Level Heart Disease Mortality
"Cleantechness" of Company Products

Cleantech Topics

- Disaster Prevention
- Energy Transportation
- Energy Production
- Energy Efficiency
- Mobility
- Air and Environment
Age and Gender of "Anonymous" Users

Goal: Predict age (18-24, 25-34, 35-49, 50+) and gender (male/female) of Twitter users

Results PAN 2015:

Age: 86%
Gender: 84%

Source: Rangel et al., 2015: Overview of the 3rd Author Profiling Task at PAN 2015
Talk in Short!

Sentiment Analysis

- approx. 70% F1-score
- the more data – the better
- has important application
Thanks!!

This presentation is based on joint work with:

• Aurelien Lucchi, ETH
• Dominic Egger, ZHAW
• Fatih Uzdilli, ZHAW
• Jan Deriu, ZHAW
• Leon Derczynski, Univ. of Sheffield
• Martin Jaggi, EPFL
• Maurice Gonzenbach, ZHAW
• Valeria de Luca, ETH

Mark Cieliebak

Zurich University of Applied Sciences (ZHAW)
Email: ciel@zhaw.ch, Website: www.zhaw.ch/~ciel