Arbeitsgruppe
Medieninformatik

Gastvortrag Datamining: Twitter Sentiment Analysis

Datenbanksysteme
Sommersemester 2015

Nils Haldenwang, M.Sc.

Datamining

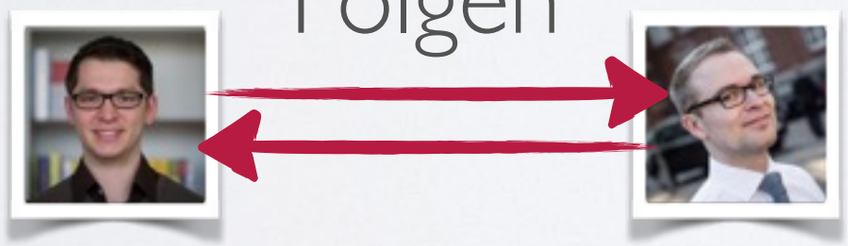
Extraktion von nützlichen Informationen aus Daten

Twitter

www.twitter.com

- Microblogging (180 Zeichen)
- 300 Millionen Nutzer
- > 500 Millionen Nachrichten/Tag
- Soziales Netzwerk

Folgen



Timeline

Anatomie von Tweets

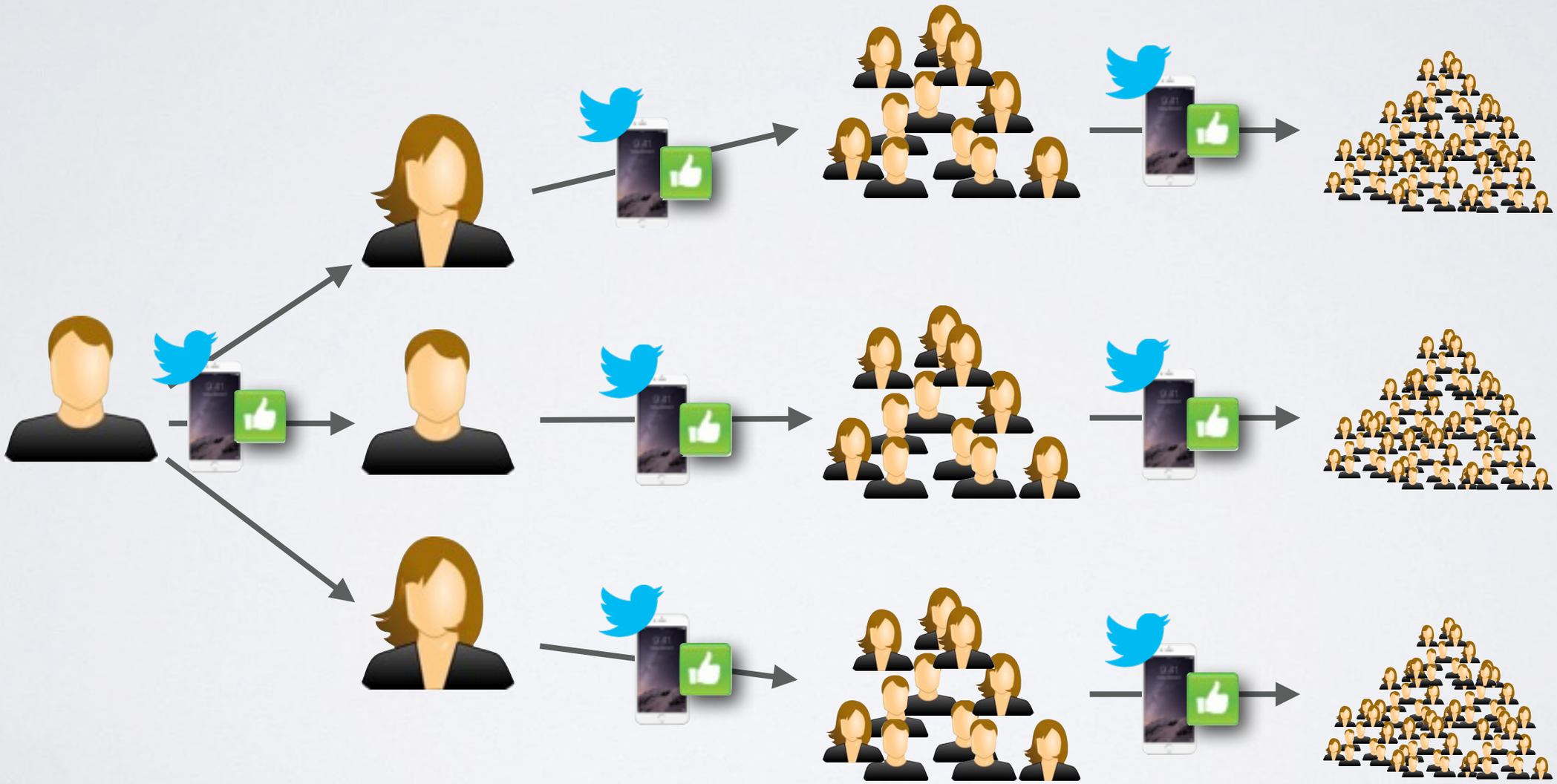
Retweet

Hashtag

Link

Twitter

= Electronic Word of Mouth



Sentiment Analysis

Xavier L @LeXavTwit · Sep 11

iPhone 5S vs **6** vs **6+**. **6+** way **too big**! **Too** bad **iPhone 6** isn't 1080, so much space lost! Touch ID should be in **display**.

View photo

Camila Salazar @camilaa_salazar · 11h

Haven't charged my phone since 8 AM & it hasn't died aka **iPhone 6** battery **is great**

4 5

Anwendungen

Marktforschung

Informationen

Krisenerkennung

Politik

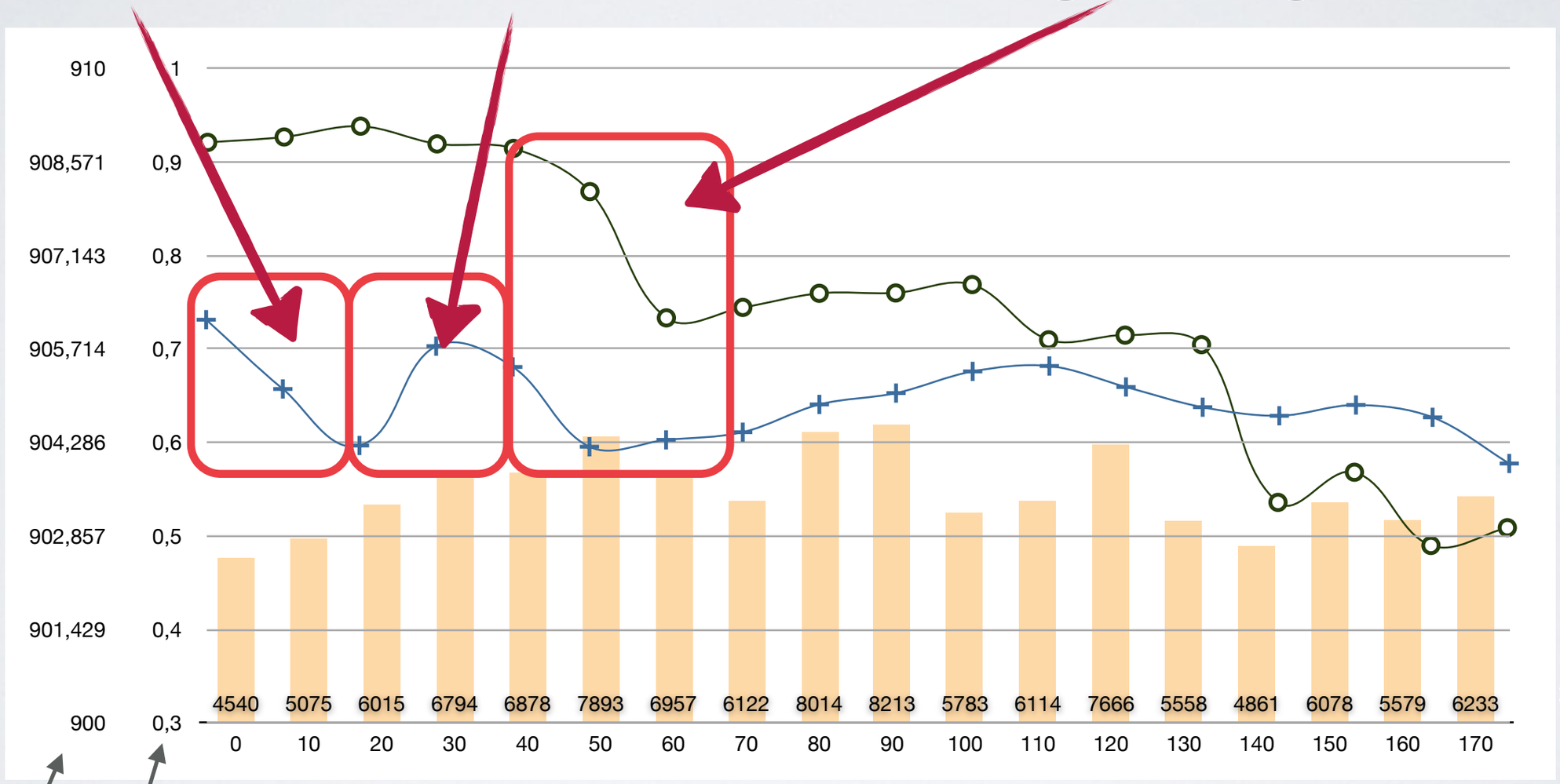
Finanzmarkt

Google IO 2013

Vorfreude

Android Studio

Fehlgeschlagene Demo



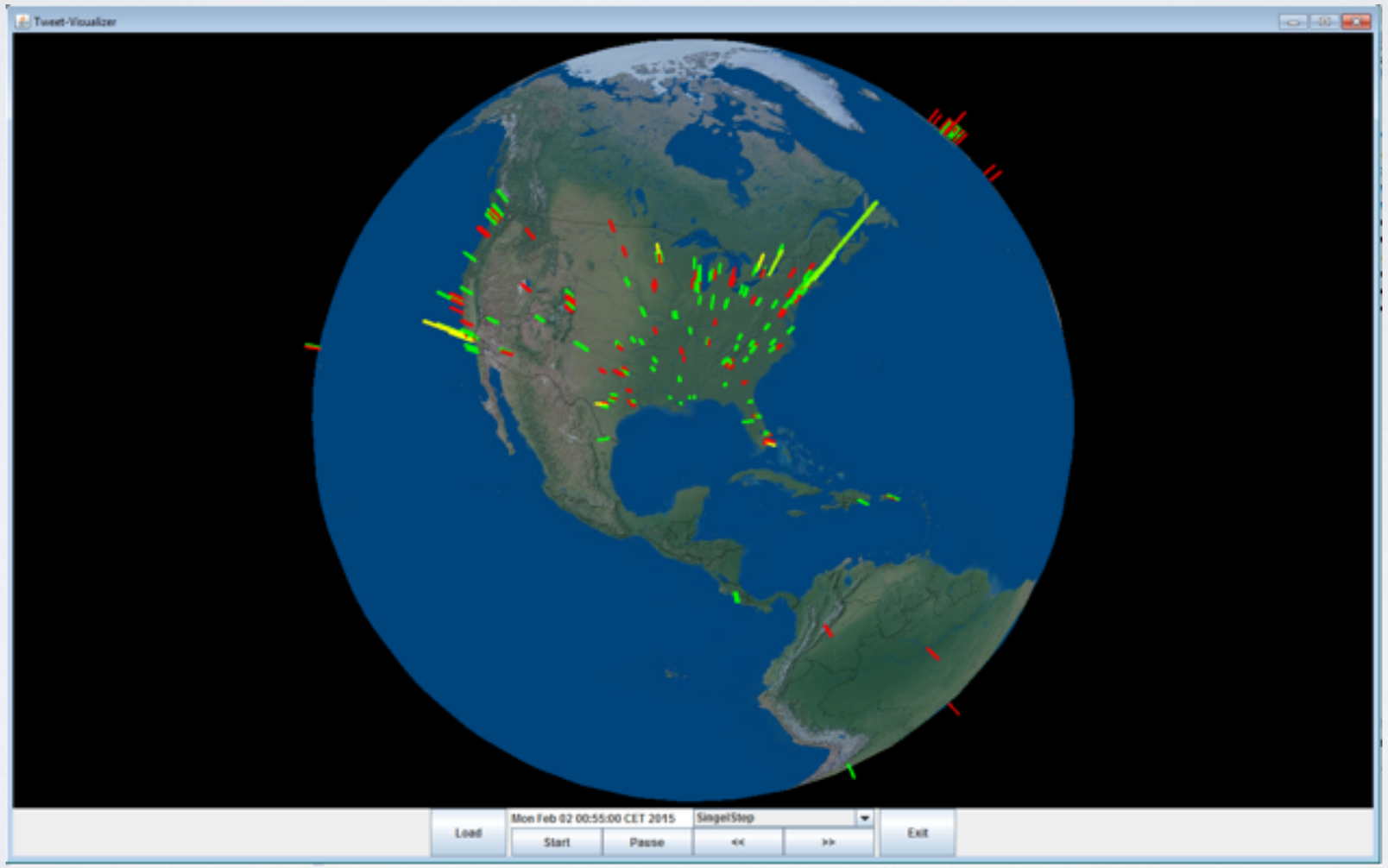
- Aktienkurs
- Stimmungs-Score

$$\frac{N(positive\_tweets) - N(negative\_tweets)}{N(positive\_tweets) + N(negative\_tweets)}$$

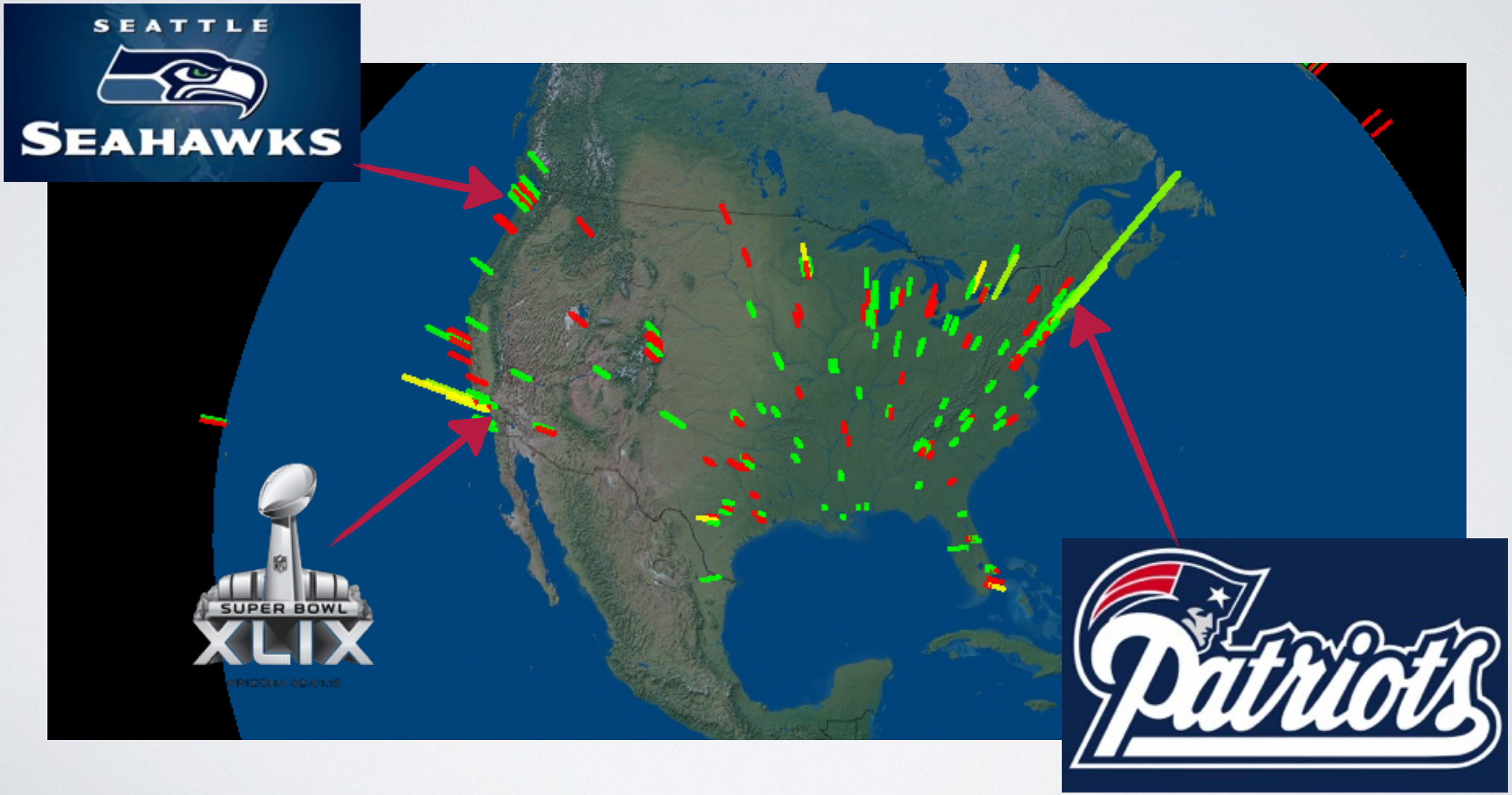
Neubauer, Nicolas. *Semantik und Sentiment: Konzepte, Verfahren und Anwendungen von Text-Mining*. Diss. 2014.

3D Stimmung

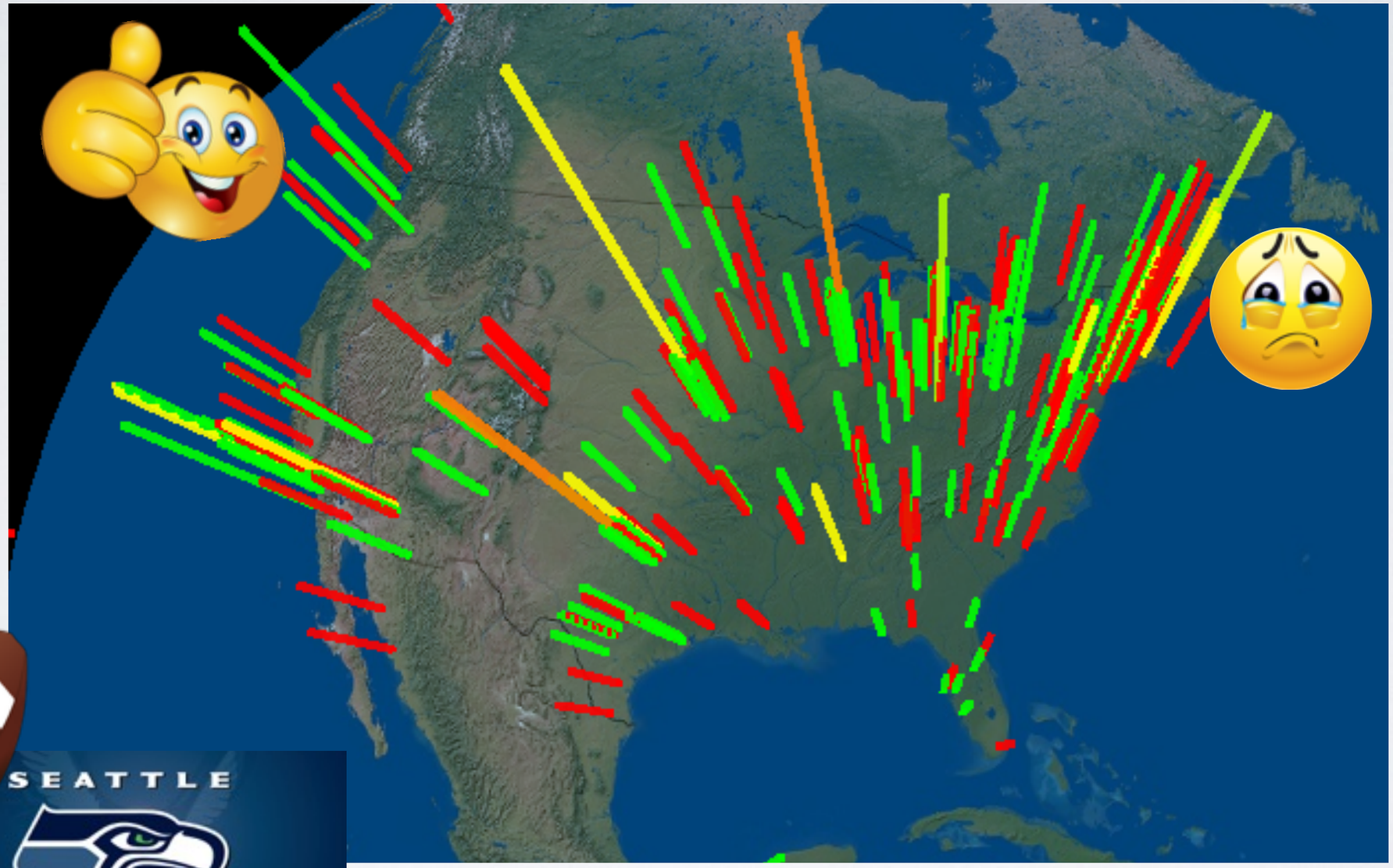
Masterarbeit Florian Dölker



Superbowl 2015



Superbowl 2015



SEATTLE
SEAHAWKS

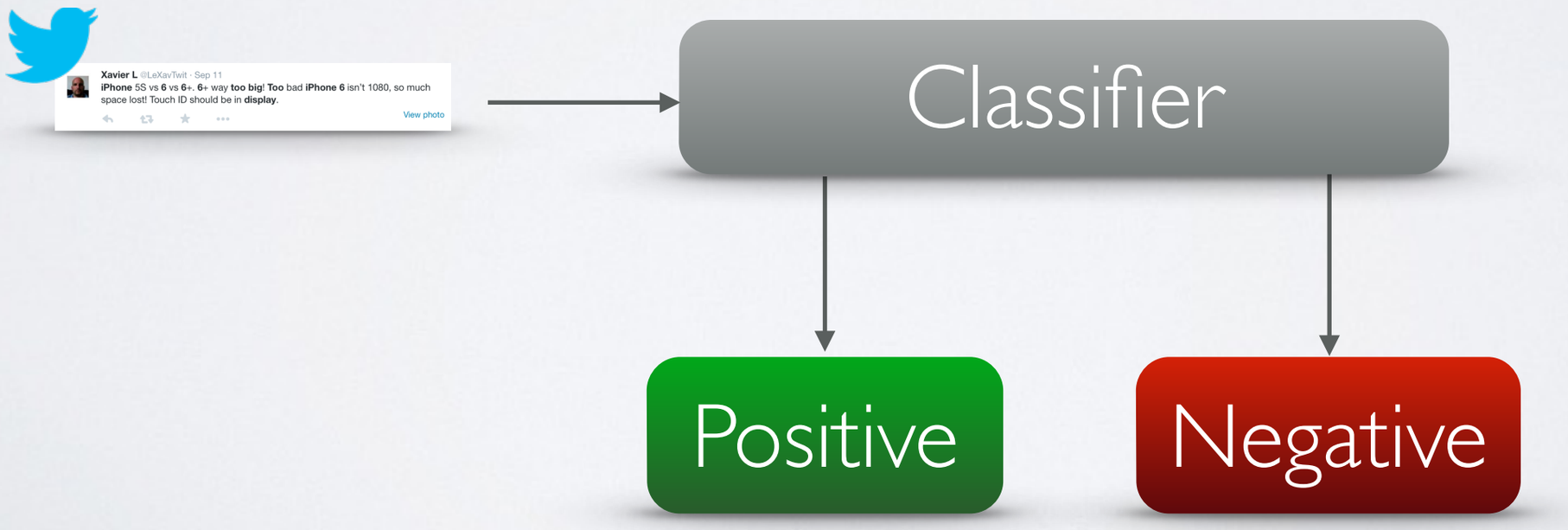
Touchdown!

Arbeitsgruppe
Medieninformatik

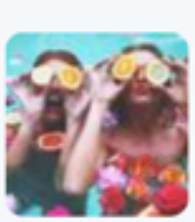
Twitter Sentiment Analysis als Klassifikationsproblem

Klassifikation

Einordnung von Objekten in gegebene Klassen anhand der Objekteigenschaften.



Dynamische Sprache



alex gunnell @alexgunnell8 · 24 Apr 2013

@_salweetah oh yeah guurrrrllll. but **dat boy** a lady killa **fo sho**. hahahahhahaha

View conversation

=

"I think I can agree to that my lady. This adolescent surely is a casanova! Hilarious!"

Klassifikationsverfahren

Händisch erstelle Classifier

z.B. nachschlagen von Begriffen in einem Sentiment-Lexikon, Erstellung einer Menge von Regeln

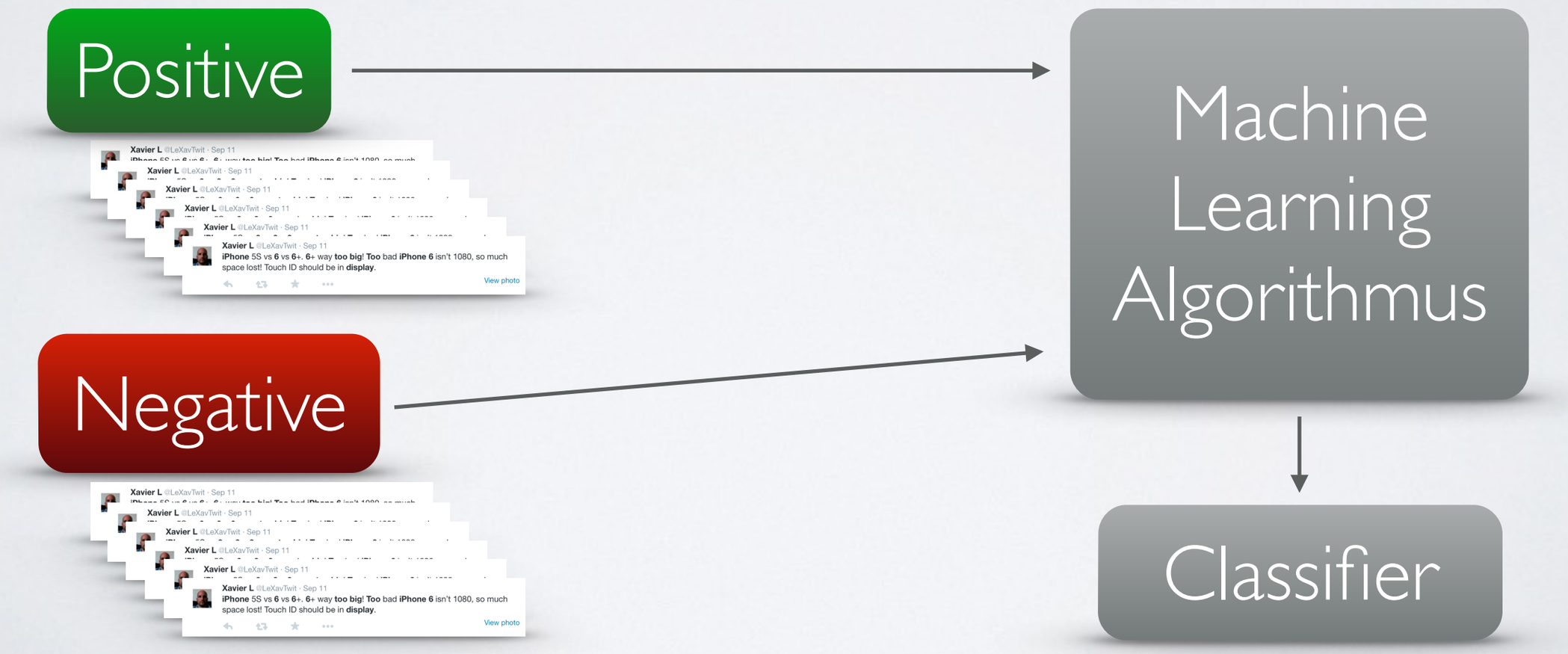
Machine Learning

Lernen ein Modell zur Klassifikation aus gegebenen Daten

Supervised Learning

Lerne aus Trainingsdaten mit bekannten Kategorien

Trainingsdaten



Lerndaten

Standardverfahren: Händisch erstellte Labels

Idee: Emoticons als Noisy Labels

"Schönes Wetter heute! :-)"

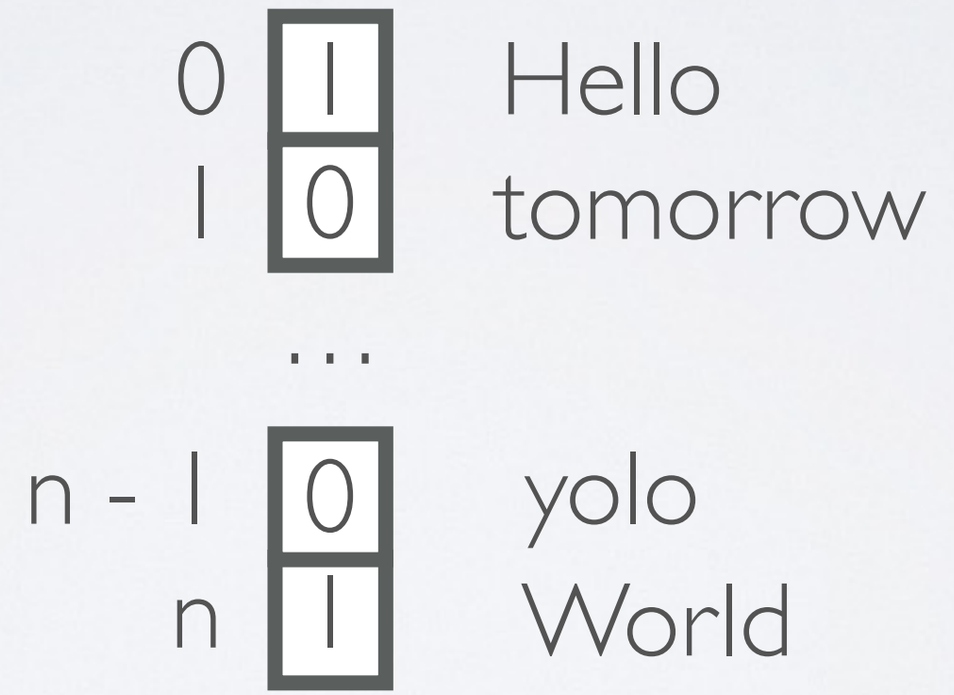
"Hausaufgaben machen. :-("

Arbeitsgruppe
Medieninformatik

Text zu Vektor

Bag-of-Words

"Hello World!"



Dimension n: Größe des Wortschatzes

Text zu Vektor

Preprocessing

Ziel: Informationsgehalt erhöhen und Dimension reduzieren

Spelling Correction

love
looove → loove
looooooove → loove

Negation Annotation

this is not cool
↓
this is not !cool

Lemmatization

was → be
am → be
have been → be

Stop Word Removal

this is hot stuff
↓
hot stuff

Acronym Expansion

lol → laughing out loud

Text zu Vektor

N-Gram Features

"What did the fox say?"

n=1:

what
did
the
fox
say

n=2:

what did
did the
the fox
fox say

n=3:

what did the
did the fox
the fox say

Vorteil:
Erfasst Struktur

Nachteil:
Erhöht Dimension

6 Mio. 35 Mio. 80 Mio.

Handcrafted Features

0 8 number of upper case letters

1 ! number of exclamation marks

...

n 42 text length

Dimension: Anzahl Features

Word Embeddings

Dimension: 50-200, reelle Zahlen

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.
[2] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013.
[3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proceedings of NAACL HLT, 2013.
[4] https://code.google.com/p/word2vec/

Machine Learning Algorithmen

Naive Bayes Classifier

probabilistic classifier

Klasse Features

$$p(C|F_1, \ldots, F_n)$$

Annahme: Die Features sind unabhängig voneinander und normalverteilt

$$p(C|F_1, \ldots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C)$$

Konstanten

Naive Bayes Classifier

Maximum-Likelihood-Estimation:

$$p(F_i|C) = \frac{f_{ic}}{f_{itotal}}$$

Vorkommen in C

Vorkommen gesamt

$$\text{classify}(f_1, \ldots, f_n) = \text{argmax}_c \prod_{i=1}^{n} \left(\frac{f_{ic}}{f_{itotal}} \right)$$

Naive Bayes Classifier

"Das Leben ist schön!"

p'(positiv | das, leben, ist, schön) =

0.5 * 0.76 * 0.5 * 0.95 = 0.1805

p'(negativ | das, leben, ist, schön) =

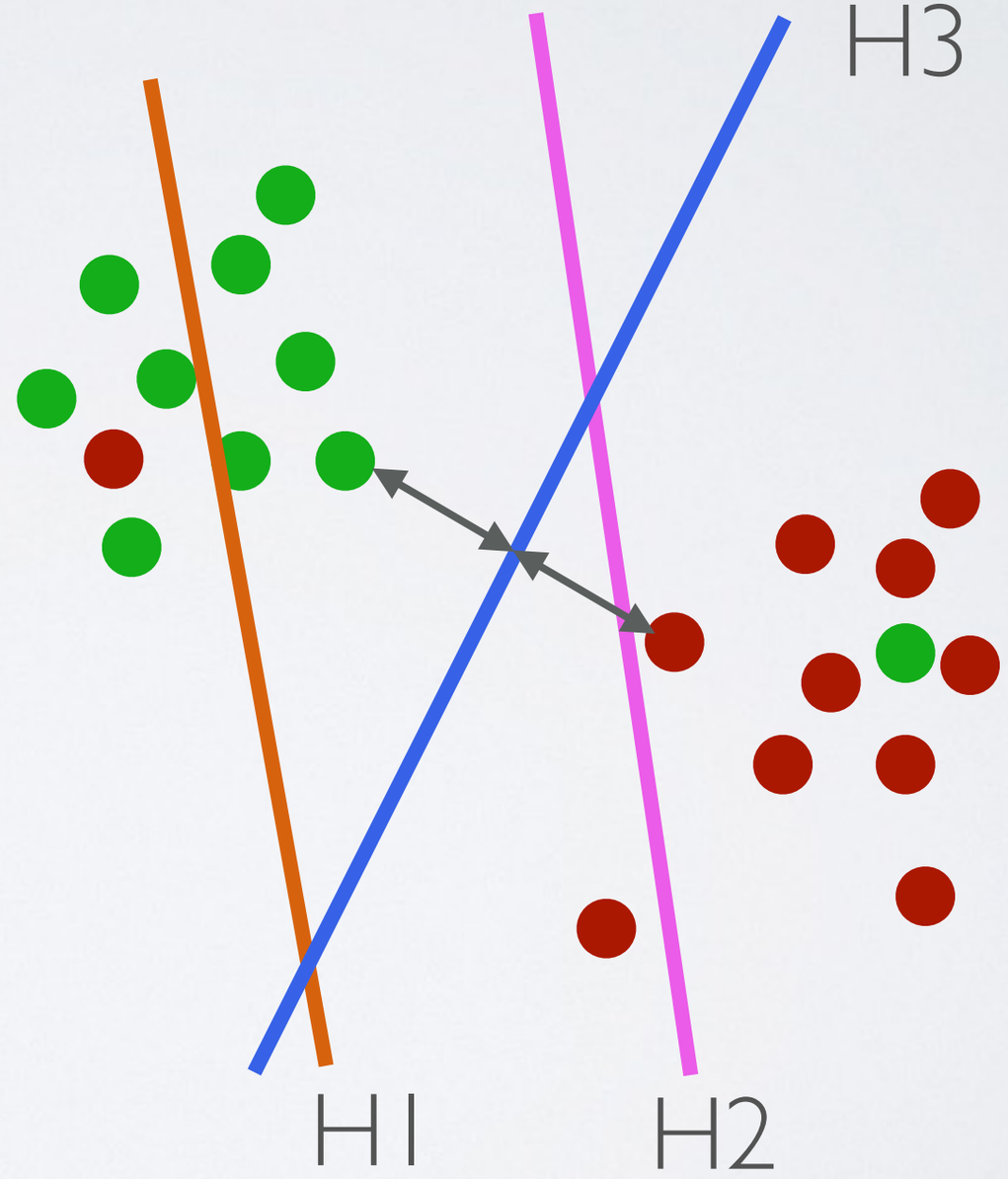
0.5 * 0.24 * 0.5 * 0.05 = 0.003

Support Vector Machines

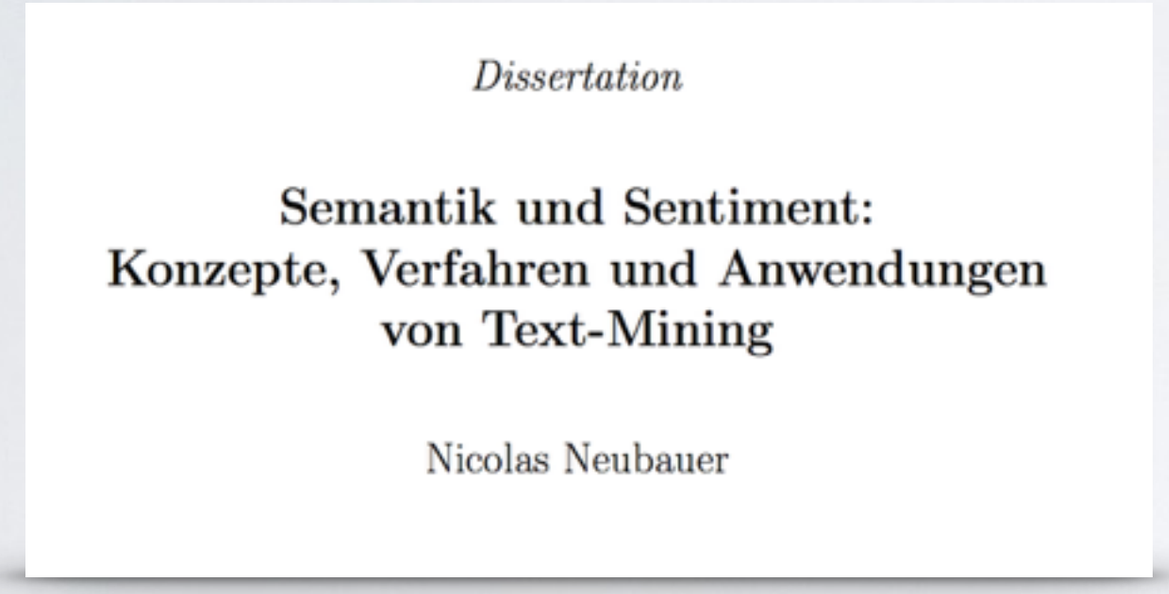
non-probabilistic binary linear classifier

Trenne Klassen durch Hyperebene

Klassifikation unbekannten Datums: Seite der Ebene



Zum weiterlesen



Dissertation

Semantik und Sentiment:
Konzepte, Verfahren und Anwendungen von Text-Mining

Nicolas Neubauer

2. Teil bietet Übersicht über Twitter Sentiment Analysis, erhältlich in der Bibliothek als PDF

ML Algorithmen

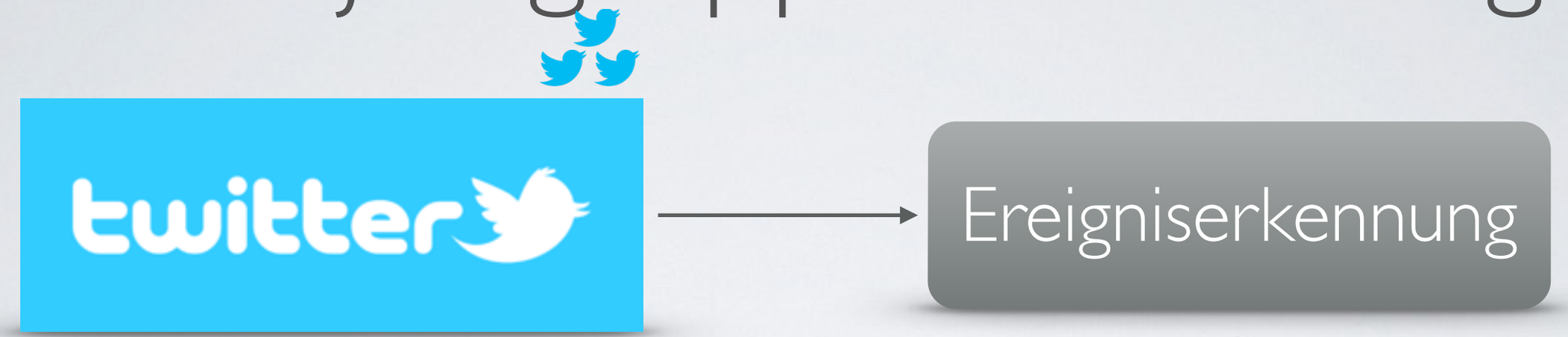
Demo

Naive Bayes Classifier
1-Gram (UniGram) Features
kein Preprocessing
ca. 12 Millionen Trainingstweets

Arbeitsgruppe
Medieninformatik

Reliable Twitter Sentiment Analysis

Projektgruppe Datamining



Mehr als nur positive/negative

uncertain

John Doe @whoami, June 30

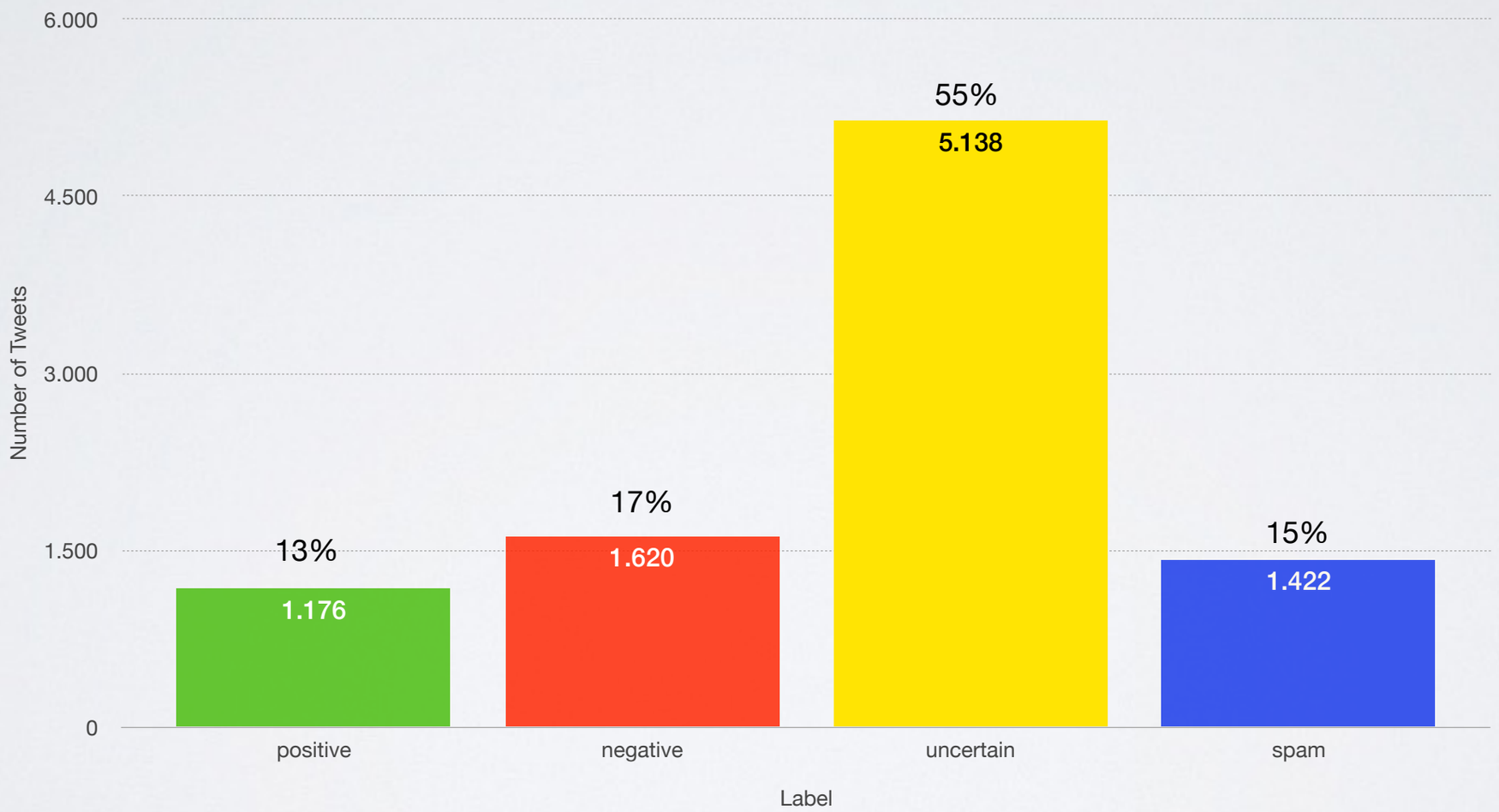
You know when you been waiting for someone to say something to you , and you got all the shit you gonna say to them prepared. Lol

spam

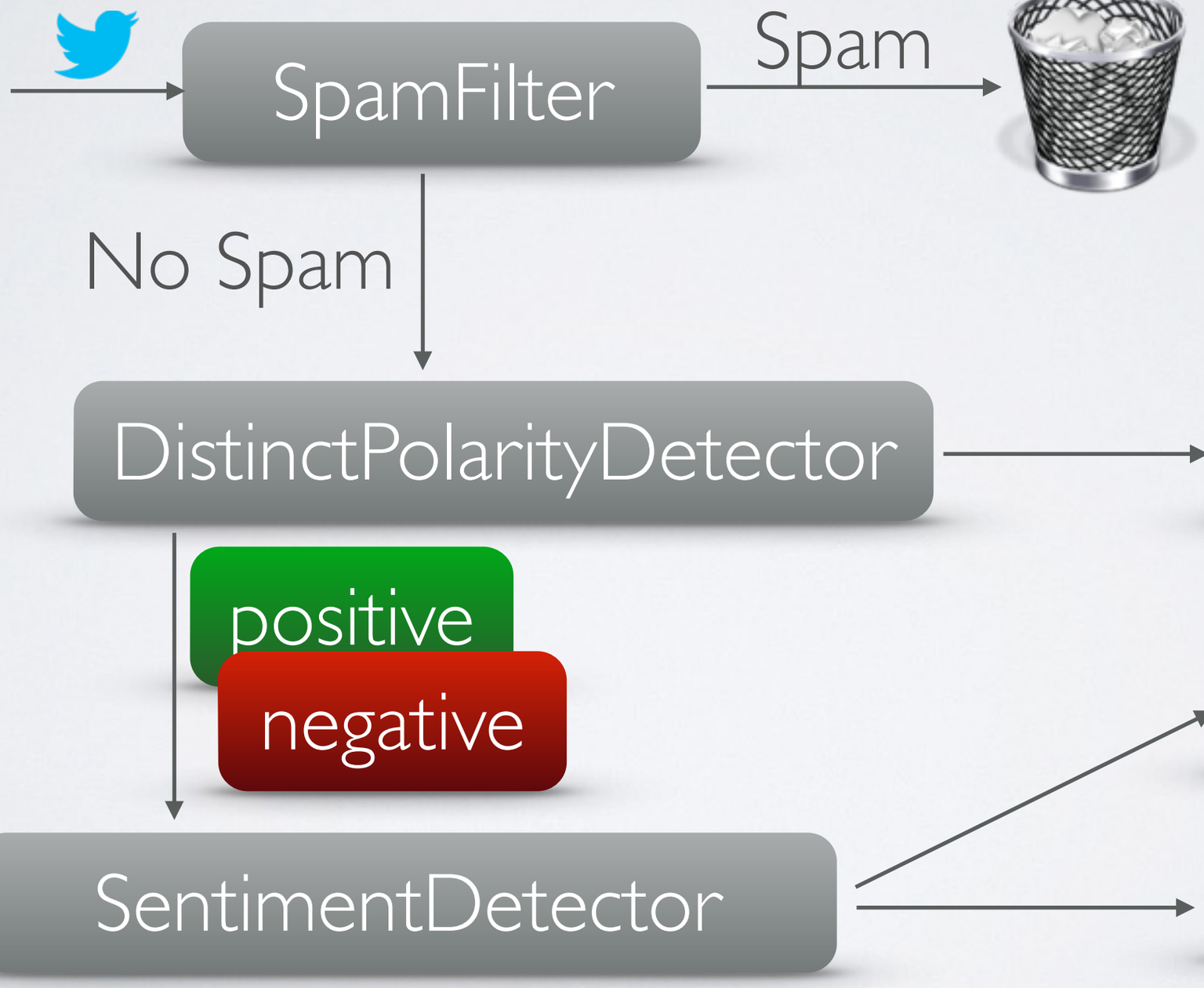
Jane Doe @soundso, June 30

I made $58.08 this week by taking 11 surveys! They only took 15 mins each :) Look here http://t.co/ejNRrtsQ

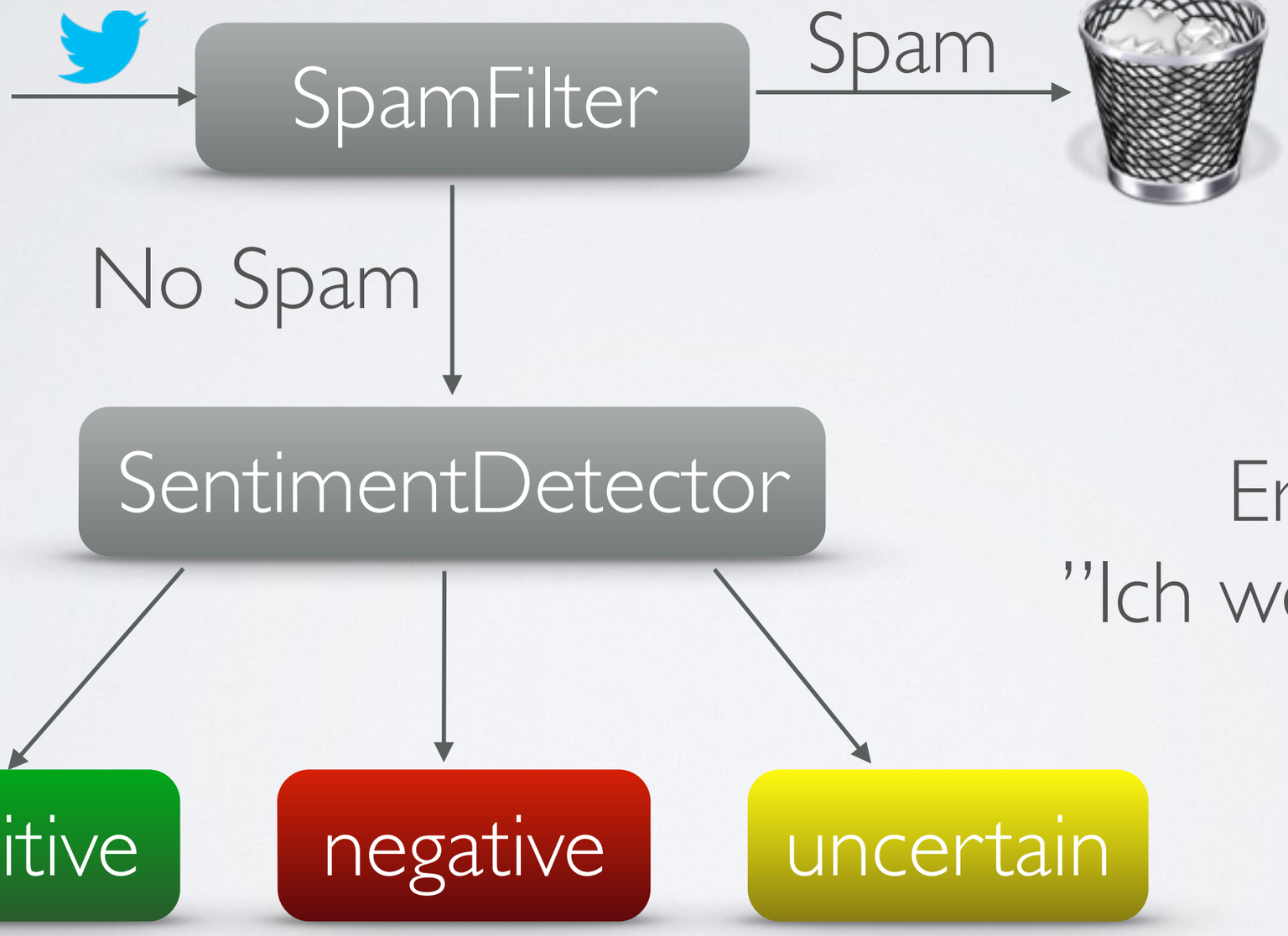
Der Public Twitter Stream



Lösungsansatz I



Lösungsansatz II



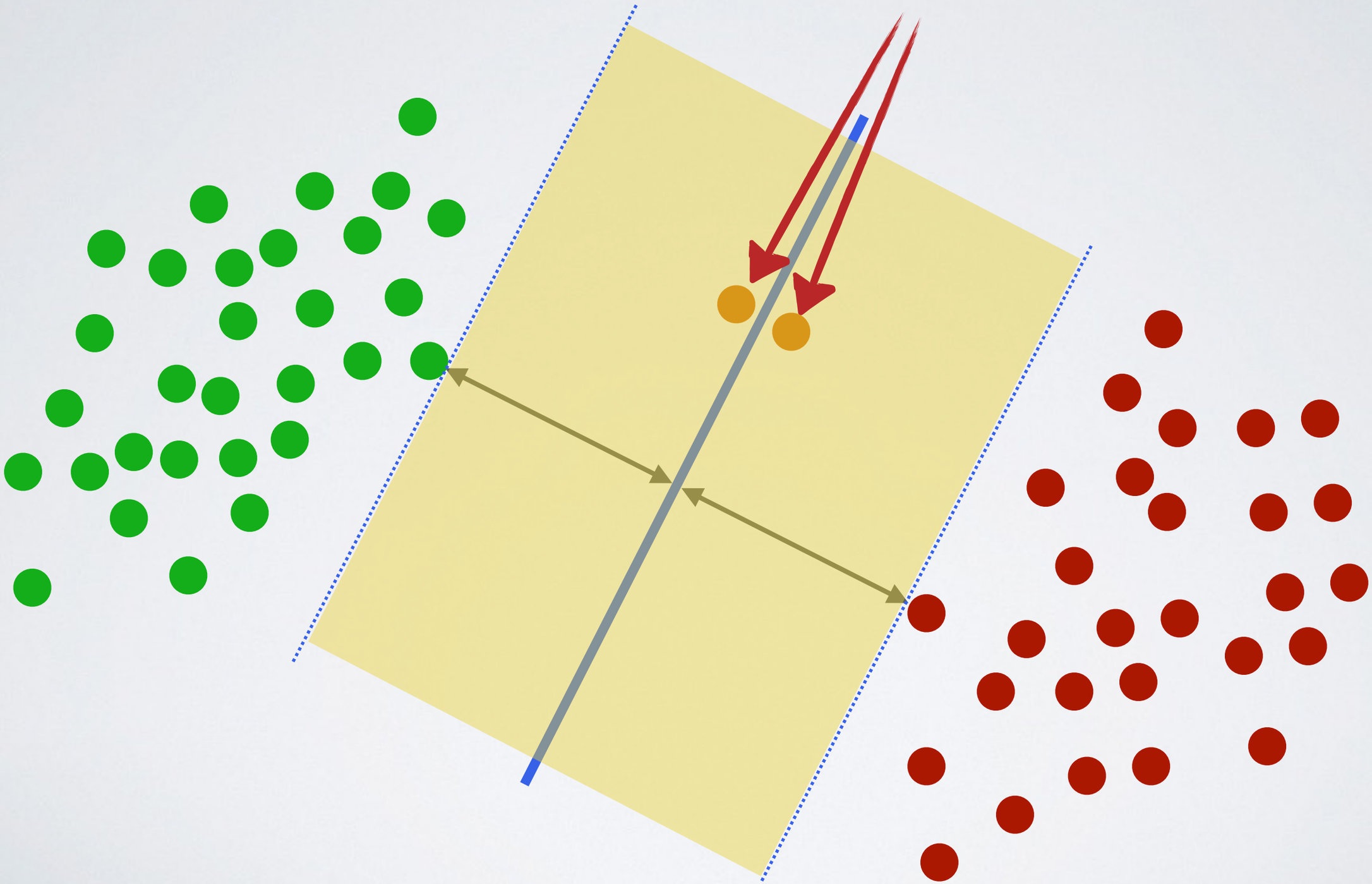
Erlauben:
"Ich weiß es nicht!"

NBC: "Ich weiß nicht!''

Betrachte Unterschied der Wahrscheinlichkeiten für die beiden Klassen

$$confidence(P_{pos}, P_{neg}) = \left| \frac{P_{pos} - P_{neg}}{P_{pos} + P_{neg}} \right|$$

SVM: "Ich weiß es nicht!"



Güte von Lösungen messen

Notwendig: Valide Testdaten

müssen händisch erstellt werden

Labeled tweets: 1510 (Rank 6 of 27, next in 641 tweets)

RAY BAN 2140, RB2140 100251 TOP BLACK ON ORANGE PLASTIC CRYSTAL BROWN GRADIENT LENS ORIGINAL... http://t.co/oMnSISSB

positive negative unclear spam

Instructions hide

Dataset Instructions :

- - - - - - - - - - - - - - - - - -

In diesem Datenset soll zunächst zwischen Marketing/Spam und Tweets von richtigen Menschen unterschieden werden. Wenn du der Meinung bist, dass der Tweet Spam/Werbung/Marketing ist, dann wähle das Label "spam". Wenn du der Meinung bist ein Mensch hat die Nachricht geschrieben und sie hat Inhalt, der kein Marketing/Werbung/Spam ist, dann wähle zwischen den Labels "positive", "negative", "unclear". Siehe unten für eine genauere Beschreibung der einzelnen Labels mit Beispielen.