Sentiment analysis of Twitter microblogging posts

Jasmina Smailović

Jožef Stefan Institute

Department of Knowledge Technologies

Introduction

- Popularity of microblogging services
- Twitter microblogging posts are short (up to 140 characters)
- Known as *tweets*
- Around 6,000 tweets are posted every second!

- In order to analyze opinions in tweets, we apply **sentiment analysis**

  The movie was fabulous! 

  The movie was horrible! 

Outline

- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

Outline

- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

The Train Dataset

- 1,600,000 labeled tweets
- Positive and negative emoticons as labels
- Origin: Go et al. (2009)

Examples:

+ Goodnight everyoneeee :) Love yall
+ I have a good feeling about today ;)
+ ooo the ice cream van is here... yaaaaaay :D

…

- I hate when I have to call and wake people up :(
- I don't have any chalk! :-/ MY CHALKBOARD IS USELESS
- UGHHHHHHHHHHHHHH.. life is NOT good all the time!!!!!! ;(

…

The Test Dataset

- 498 hand-labeled tweets
- Tweets belong to different domains
- 182 positive, 177 negative, and 139 neutral tweets
- Origin: Go et al. (2009)

Outline

- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

Sentiment Analysis Approaches

- Machine Learning
- Lexicon-based
- Linguistic approach

Sentiment Analysis Algorithm Selection

The first experiment

- Test dataset: 177 negative and 182 positive hand-labeled tweets
- The machine learning approach:
  - The linear SVM (SVM^{perf}), Naive Bayes, and k-Nearest Neighbors (the LATINO library)
  - Train dataset: 1,600,000 smiley-labeled tweets
- The lexicon-based approach:
  - The opinion lexicon (2,006 positive and 4,783 negative words) (Hu & Liu, 2004; Liu et al., 2005)

Accuracy on the test set			
SVM	NB	K-NN	Lexicon
79.11%	75.21%	72.98%	73.54%

Sentiment Analysis Algorithm Selection

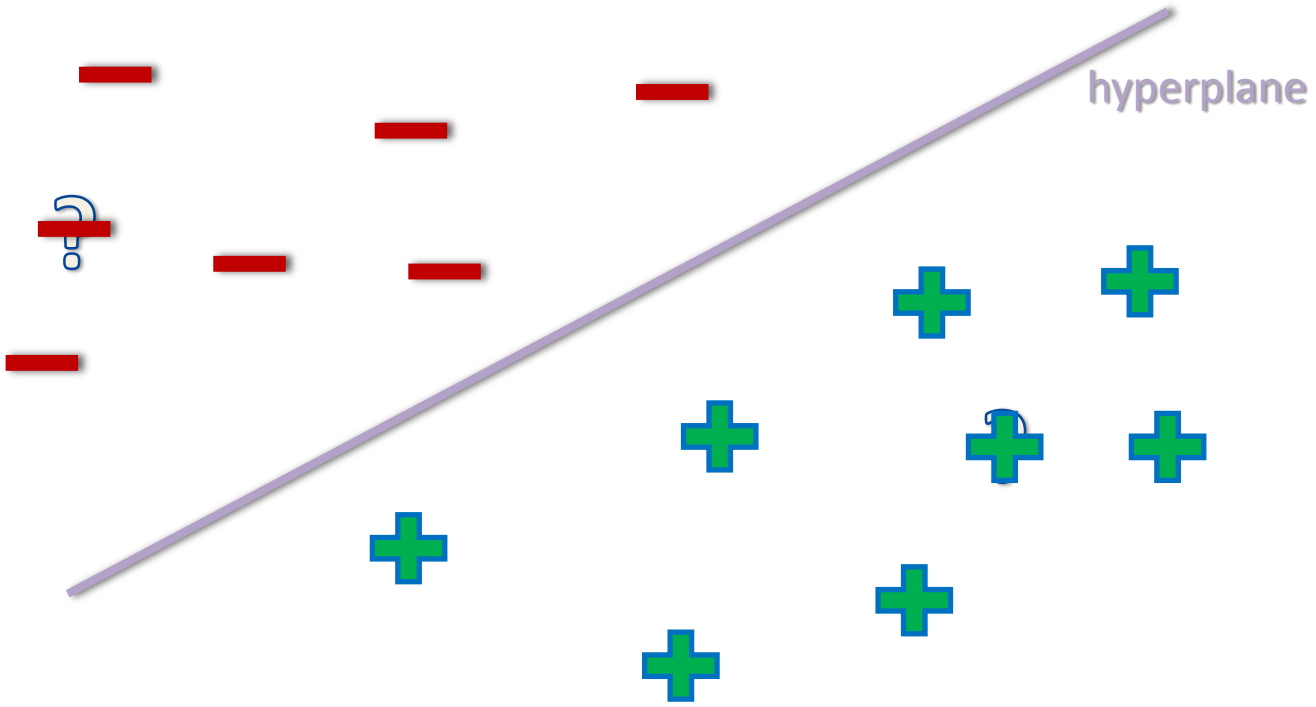
The second experiment

- Stratified ten-fold cross-validation on 1,600,000 smiley-labeled tweets
- The machine learning algorithms

10-fold cross-validation		
SVM	NB	K-NN
78.55%	75.84%	slow

- The SVM approach in used the rest of our analyses

Linear Support Vector Machine (SVM)



Outline

- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

Data preprocessing

- Unique phrases, slang, grammatical and spelling mistakes in Twitter posts

  @jenny I am with my Sisterrrrrrr and we are buying $aapl stocks #happy !

- Twitter-specific and standard preprocessing

Twitter-specific preprocessing

- Usernames *@TwitterUser → atttTwitterUser*
- Stock Symbols *$GOOG → stockGOOG*
- Usage of Web links *www.abc.com → URL*
- Hashtags *#bowling → hashbowling*
- Exclamation and question marks *(e.g., replacing ?!??!!? by the MULTIMIX token)*
- Letter repetition *goooooooooood → goood*
- Negations *not, isn't, aren't,… → NEGATION*

Standard preprocessing (1)

- Text tokenization
  - Regex
- *@jenny we are buying $aapl stocks #happy ! https://www.apple.com*
- Tokens: *<"@", "jenny", "we", "are", "buying", "$", "aapl", "stocks", "#", "happy", "!", "https", "://", "www", ".", "apple", ".", "com">*
  - Simple
- *@jenny we are buying $aapl stocks #happy ! https://www.apple.com*
- Tokens: *<"jenny", "we", "are", "buying", "aapl", "stocks", "happy", "https", "www", "apple", "com">*

Standard preprocessing (2)

- Stemming birds → *bird*
- *n*-gram construction

I drink coffee → *<i, i drink,drink, drink coffe, coffe>*

- Testing stop word removal (*a, the, and, …)*
- The condition that a given term has to appear at least twice in the entire corpus
- Constructing Term Frequency feature vectors
- A part-of-speech (POS) tagger was not used

ID	Web links	Hashtags	Ex. and q. marks	Letter repet.	Neg.	Stop words	Avg. accuracy ± std. dev.	Avg. F-measure ± std. dev.
1		X		X			81.23% ± 0.16%	0.8143 ± 0.0046
2	X		X				81.07% ± 0.33%	0.8127 ± 0.0049
3			X		X		81.09% ± 0.20%	0.8125 ± 0.0067
4	X						81.26% ± 0.21%	0.8123 ± 0.0047
5					X		81.18% ± 0.21%	0.8121 ± 0.0041
6	X			X			81.24% ± 0.19%	0.8121 ± 0.0047
7	X	X	X	X			81.25% ± 0.31%	0.8116 ± 0.0073
8	X	X		X	X		81.23% ± 0.33%	0.8113 ± 0.0064
9		X	X	X	X		81.16% ± 0.24%	0.8110 ± 0.0080
10	X			X	X		81.24% ± 0.22%	0.8110 ± 0.0049
11	X		X	X			81.12% ± 0.20%	0.8109 ± 0.0046
12	X	X	X	X	X		81.15% ± 0.20%	0.8109 ± 0.0042
13	X				X		81.08% ± 0.20%	0.8109 ± 0.0060
14	X	X	X		X		81.19% ± 0.19%	0.8108 ± 0.0044
15	X		X	X	X		81.21% ± 0.26%	0.8106 ± 0.0065
16		X	X	X			81.14% ± 0.20%	0.8105 ± 0.0065
17		X	X		X		81.19% ± 0.16%	0.8104 ± 0.0048
18		X					81.04% ± 0.30%	0.8103 ± 0.0077
19							81.13% ± 0.24%	0.8100 ± 0.0056
20		X	X				81.13% ± 0.15%	0.8099 ± 0.0048
21				X			81.15% ± 0.32%	0.8099 ± 0.0078
22			X	X	X		81.12% ± 0.34%	0.8096 ± 0.0080
23	X	X	X				81.04% ± 0.20%	0.8093 ± 0.0064
24			X				81.15% ± 0.25%	0.8089 ± 0.0062
25				X	X		81.10% ± 0.20%	0.8086 ± 0.0057
26	X	X		X			81.15% ± 0.17%	0.8086 ± 0.0047
27	X		X		X		81.09% ± 0.28%	0.8086 ± 0.0054
28		X		X	X		81.08% ± 0.20%	0.8079 ± 0.0051
29	X	X			X		81.10% ± 0.24%	0.8077 ± 0.0062
30			X	X			81.26% ± 0.22%	0.8076 ± 0.0033
31	X	X					81.10% ± 0.21%	0.8074 ± 0.0061
32		X			X		81.09% ± 0.30%	0.8072 ± 0.0060
33	X	X		X	X	X	79.16% ± 0.23%	0.7962 ± 0.0050
34		X			X	X	79.11% ± 0.24%	0.7942 ± 0.0045
35	X		X		X	X	79.22% ± 0.20%	0.7936 ± 0.0057
36		X	X	X	X	X	79.20% ± 0.18%	0.7931 ± 0.0067
37	X		X	X	X	X	79.22% ± 0.18%	0.7930 ± 0.0066
38	X	X	X	X	X	X	79.21% ± 0.24%	0.7930 ± 0.0072
39	X	X	X		X	X	79.21% ± 0.21%	0.7926 ± 0.0054
40				X	X	X	79.26% ± 0.23%	0.7926 ± 0.0061
41		X	X		X	X	79.20% ± 0.20%	0.7925 ± 0.0066
42			X	X	X	X	79.23% ± 0.24%	0.7923 ± 0.0038
43		X		X	X	X	79.23% ± 0.23%	0.7923 ± 0.0044
44	X			X	X	X	79.18% ± 0.22%	0.7905 ± 0.0054
45	X	X			X	X	79.21% ± 0.19%	0.7904 ± 0.0044
46			X		X	X	79.21% ± 0.21%	0.7903 ± 0.0058
47					X	X	79.17% ± 0.24%	0.7902 ± 0.0059
48	X				X	X	79.18% ± 0.20%	0.7902 ± 0.0056
49	X	X				X	78.48% ± 0.21%	0.7900 ± 0.0060
50	X					X	78.45% ± 0.22%	0.7882 ± 0.0061
51	X			X		X	78.58% ± 0.21%	0.7881 ± 0.0067
52		X		X		X	78.53% ± 0.19%	0.7878 ± 0.0066
53			X	X		X	78.62% ± 0.25%	0.7878 ± 0.0052
54	X		X	X		X	78.63% ± 0.15%	0.7877 ± 0.0060
55	X		X			X	78.53% ± 0.20%	0.7870 ± 0.0057
56		X				X	78.54% ± 0.14%	0.7868 ± 0.0052
57		X	X	X		X	78.60% ± 0.25%	0.7867 ± 0.0047
58				X		X	78.52% ± 0.17%	0.7865 ± 0.0074
59						X	78.45% ± 0.24%	0.7863 ± 0.0060
60	X	X	X	X		X	78.49% ± 0.16%	0.7860 ± 0.0053
61	X	X		X		X	78.50% ± 0.21%	0.7860 ± 0.0066
62	X	X	X			X	78.55% ± 0.21%	0.7858 ± 0.0070
63		X	X			X	78.49% ± 0.15%	0.7855 ± 0.0068
64			X			X	78.43% ± 0.20%	0.7823 ± 0.0063

Preprocessing experiments

Stratified ten-fold cross-validation on 1,600,000 smiley-labeled tweets

64 combinations

The best one:

- Avg. accuracy 81.23% ± 0.16%
- Avg. F-measure 0.8143 ± 0.0046

- 1,198,302 features
- The accuracy of 80.22% on the test dataset

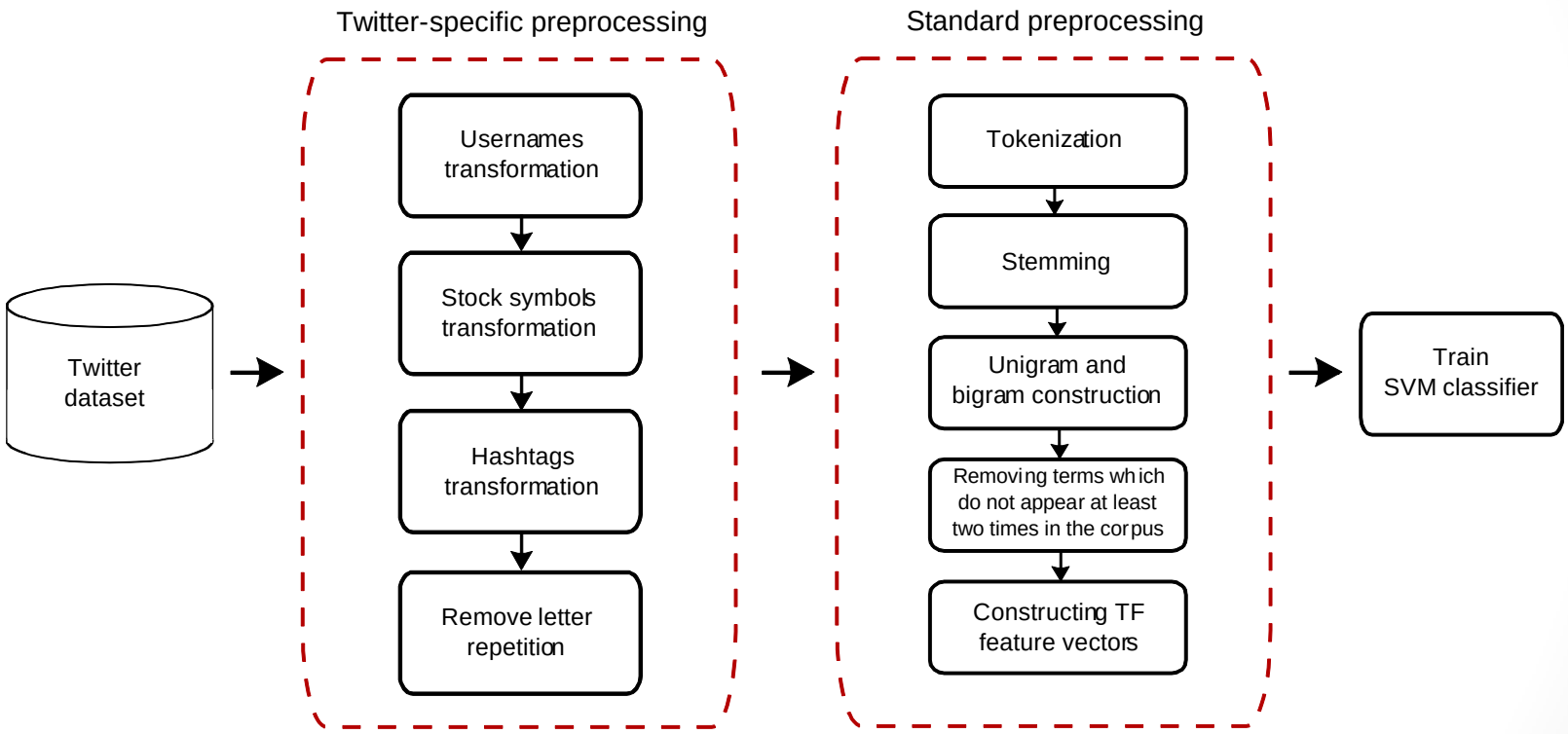
Preprocessing example

- *@jenny I am with my Sisterrrrrrr and we are buying $aapl stocks #happy !*

- *atttjenny i am with my sisterrr and we are buying stockaapl stocks hashhappy !*

- Features: *atttjenni, atttjenni i, i, i am, am, am with, with, with my, my, my sisterrr, sisterrr, sisterrr and, and, and we, we, we are, are, are buy, buy, buy stockaapl, stockaapl, stockaapl stock, stock, stock hashhappi, hashhappi, hashhappi !, !*

Proposed Preprocessing Steps



Comparison With Publicly Available Sentiment Classifiers

- Performance testing on hand-labeled tweets
  (Go et al., 2009)

Sentiment tool	Accuracy on the test set
Our approach	80.22%
AlchemyAPI	83.57%
Repustate	66.57%
Text-processing	61.00%
Sentiment140	45.96%
SentiStrength	69.92%

- Advantages of our approach:
  - Classification of much larger sets of tweets
  - Tweet preprocessing

Outline

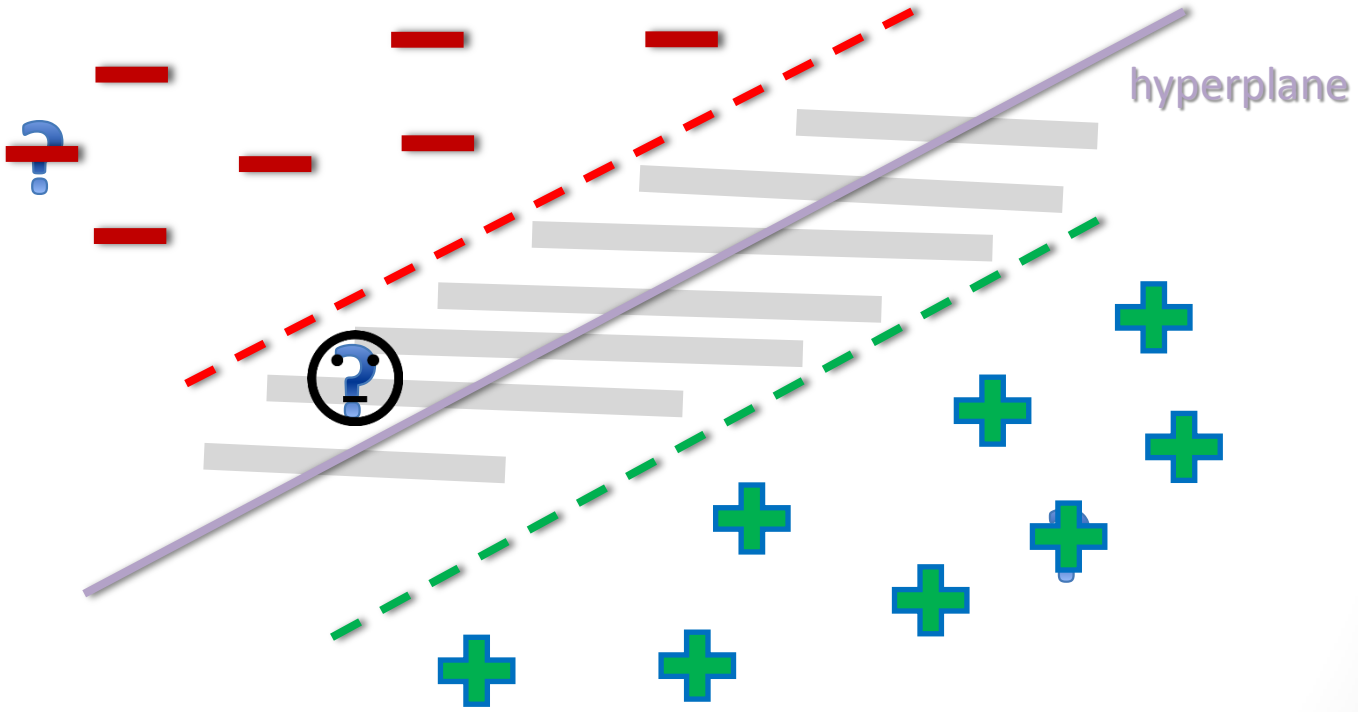
- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

The SVM Neutral Zone

- A tweet should also have the possibility of being classified as neutral or weakly opinionated

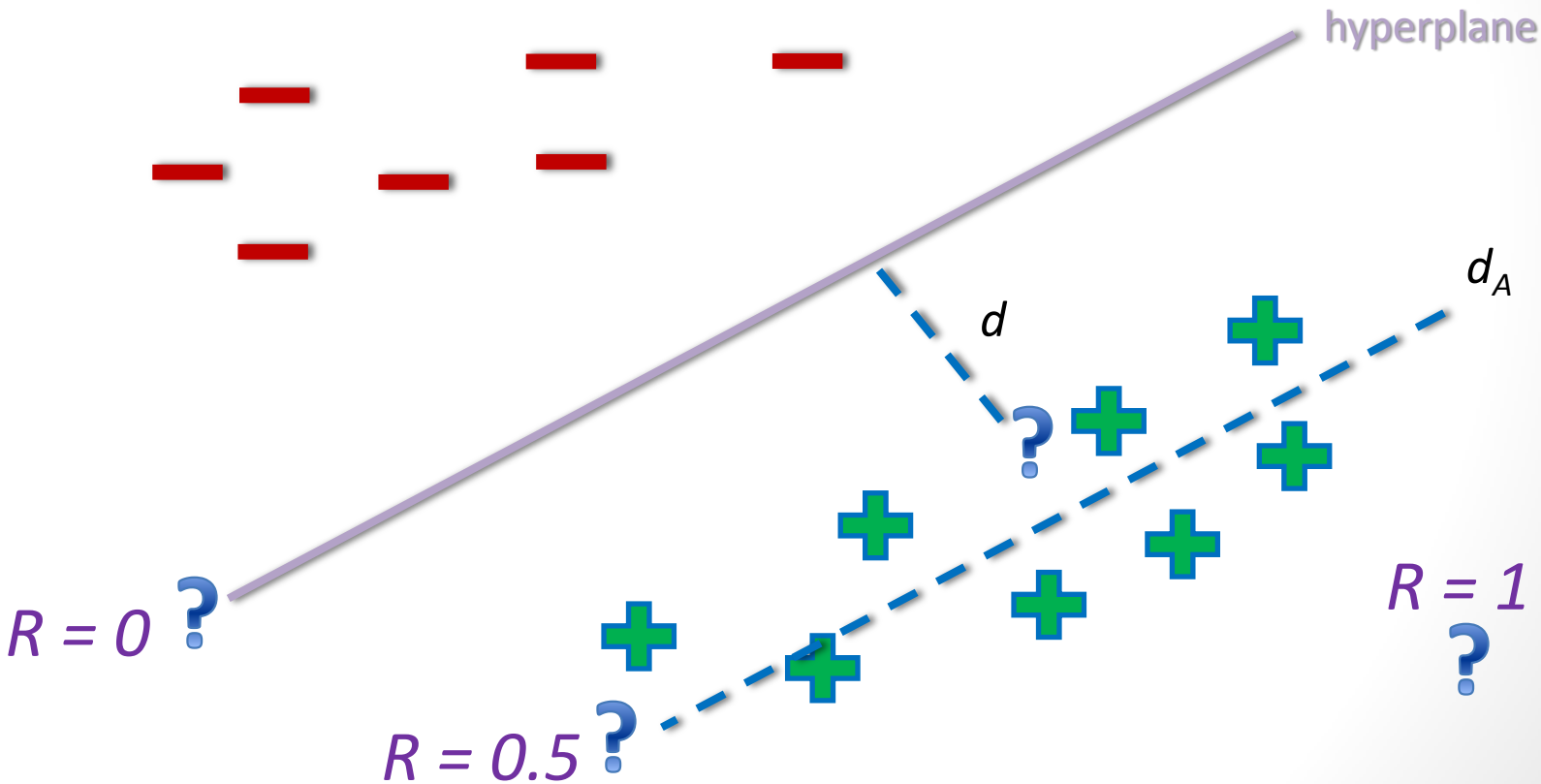
- Two ways of identifying non-opinionated tweets:
  - Fixed neutral zone
  - Relative neutral zone

Fixed Neutral Zone



Relative Neutral Zone

$$R = \begin{cases} \frac{d}{2 \times d_A} & \text{if } d < 2 \times d_A \\ 1 & \text{if } d \geq 2 \times d_A \end{cases}$$



Outline

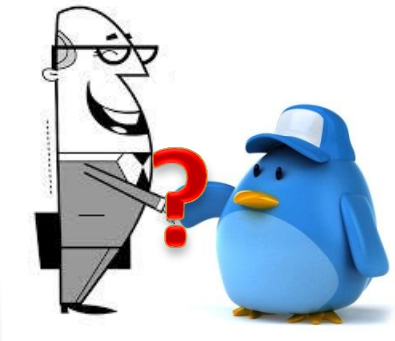
- Twitter Datasets
- Sentiment Analysis Algorithm
- Data Preprocessing
- Identifying non-opinionated tweets
- Real-world applications of the developed sentiment analysis methodology

Real-world Applications and Public Availability

- The developed sentiment analysis methodology has been applied in:
  - Financial domain
  - Political domain
  - Environmental domain
- Public Availability:
  - The ClowdFlows data mining platform
  - The PerceptionAnalytics platform

The Stock Market Application

- Investigated whether sentiment analysis of Twitter posts is a suitable data source for predicting future stock market values
- The experiments indicated that sentiment analysis of public mood derived from Twitter feeds could be used to forecast movements of individual stock prices
- The methodology was adapted to data streams



Real-time Opinion Monitoring

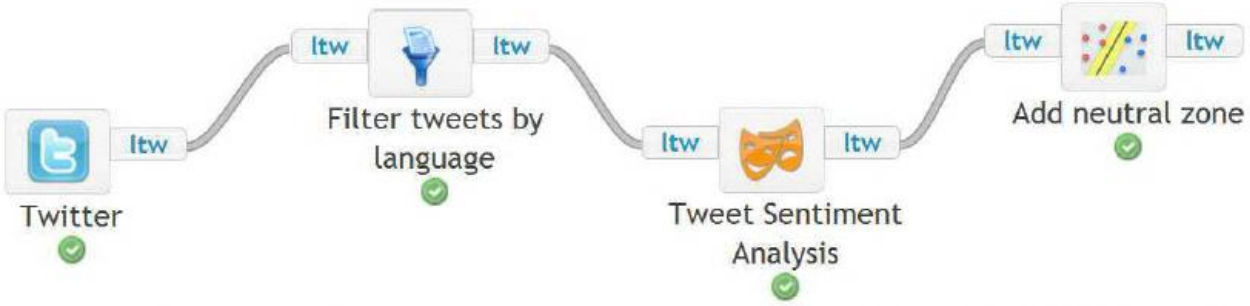
- Slovenian Presidential Elections Use Case
- Bulgarian Parliamentary Elections Use Case

Community Sentiment on Environmental Topics in Social Networks

- The developed sentiment classifier was applied on tweets discussing environmental issues
- Sentiment analysis was performed to discover the sentiment of the detected Twitter communities with respect to different topics

Implementations in the ClowdFlows Platform

- Interactive data mining platform (Kranjc et al., 2012)
- http://clowdflows.org/
- Sentiment Analysis Widget




```
Selection of "Twitter" widget Inputs:
-------
query = u'beautiful'
...
```

```
Selection of "Add neutral zone"
widget Outputs:
-------
'lang': 'en',
'reliability': 0.8552150908816,
'sentiment': 'Positive',
'text': 'RT [redacted]: oh my god shes
so beautiful http://t.co/VLMRZCvkIt',
...
```

Implementations in the PerceptionAnalytics Platform

- http://www.perceptionanalytics.net/
- A platform of a Slovenian company Gama System
- Real-time analysis
- Sentiment analysis for a number of languages: English, Slovenian, Spanish, German, Russian, Hungarian, Polish, Portuguese, Bulgarian, etc.

Bibliography

- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2014). Stream-based active learning for sentiment analysis in the financial domain. *Information Sciences*, *285*, 181–203..
- Kranjc, J., Smailović, J., Podpečan, V., Grčar, M., Žnidaršič, M., & Lavrač, N. (2014). Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform. *Information Processing & Management*. doi:http://dx.doi.org/10.1016/j.ipm.2014.04.001.
- Sluban, B., Smailović, J., Juršič, M., Mozetič, I., & Battiston, S. (2014). Community sentiment on environmental topics in social networks. In *Proceedings of the 10th International Conference on Signal Image Technology & Internet Based Systems (SITIS), 3rd International Workshop on Complex Networks and their Applications* (pp. 376–382).
- Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 77–88). Lecture Notes in Computer Science Volume 7947. Springer Berlin Heidelberg.
- Smailović, J., Grčar, M., & Žnidaršič, M. (2012). Sentiment analysis on tweets in a financial domain. In *Proceedings of 4th Jožef Stefan International Postgraduate School Students Conference* (pp. 169–175).
- Smailović, J., Žnidaršič, M., & Grčar, M. (2011). Web-based experimental platform for sentiment analysis. In *Proceedings of the 3rd International Conference on Information Society and Information Technologies (ISIT)*.

Thank you!