Faculty of Electrical Engineering and Informatics

Algorithms of Machine Learning in the Recognition of Trolls in the Online Space

SAMI 2020

Kristína Machová, Michal Porezaný and Miroslava Hrešková

Department of Cybernetics and Artificial Intelligence, Technical University of Košice

January 21-23, 2021

Antisocial Content in Online Discussions

- Positive influence of social media: connectivity, learning, access to information, etc.
- While most users tend to accept social norms, others engage in antisocial behavior negatively affecting the rest of the community
- Negative impact of social media: **trolling, fake news, hoaxes, cyber bullying, harassment, flaming, rumors, hate speech, ...**
- Some forms of regulation are in demand
- Focus on **detecting trolls** in contrast to credible contributors

Trolling

- Concept „Troll“ is an Internet slang term – an individual attempting to incite conflict by publishing **offensive**, **inflammatory**, **provocative** or **irrelevant** posts.
- Impact on a society is dangerous because of an intent:
  - to over-throw the debate, to reduce its substance,
  - to **spread of a hatred** among the discussants or causing conflicts in provocative topics
  - to **change and manipulate the opinion** of a society on some important topics (votes, politicians, health...)
  - motivation can also be financial, when trolls are members of the so-called „**troll farms**” - organized groups that are paid to publish troll’s posts, propaganda, or fake news
  - chatbots, which can pass the Turing test can have the forms of so-called “trollbots”
- Trolling is becoming more and more widespread - it is important to propose means for the automatic detection of this unhealthy web phenomena

Used Machine Learning Methods

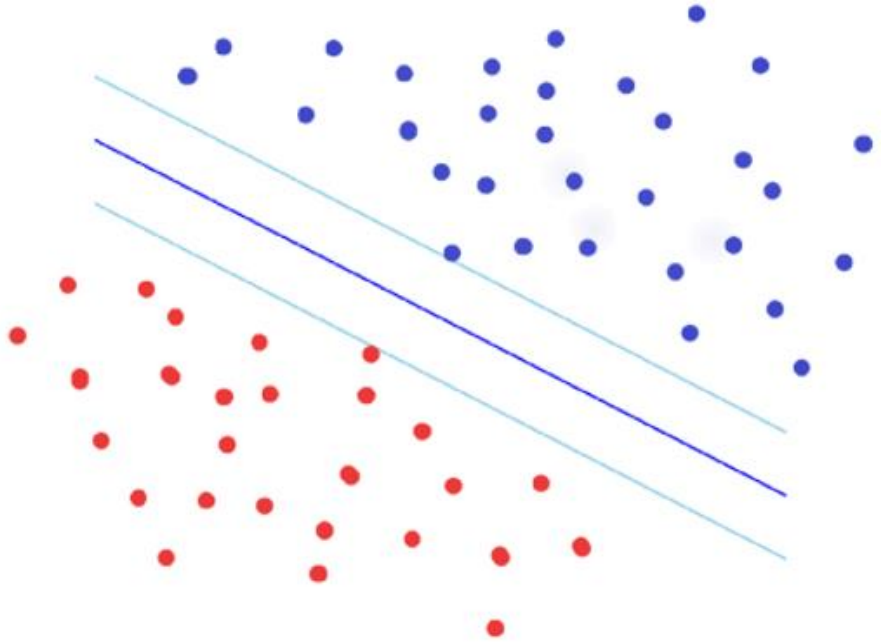
- Support Vector Machine (SVM)
  - can work with a lager attribute space than other methods (suitable for text data processing)
- Multinomial Naive Bayes Classifier (MNB)
  - suitable for attributes with discrete values
- Logistic Regression (LR)
  - only one method of regression analysis, which can be used for classification

Input for ML is a labeled training set

Output of ML is a model for final classification of an unknown contributor to troll / non-troll class, for recognition of trolls in online communities

Support Vector Machine

- SVM separates the sample space into two or more classes with the **widest margin possible**
- Originally a linear classifier
- Perform non-linear classification and based on a kernel function
- Kernel is a method which maps features into a higher dimensional space
- Only nearest points to the separating hyper-plane determine margin - they are called support vectors
- To **maximize the margin** is the primary problem of SVM



Multinomial Naive Bayes Classifier

- The probabilistic classifier based on Bayes' theorem:

$$P(A/B) = \frac{P(A)P(B/A)}{P(B)}$$

- *P(B)* can be an estimated constant β^{-1} calculated from the dataset
- Naive Bayes Classifier

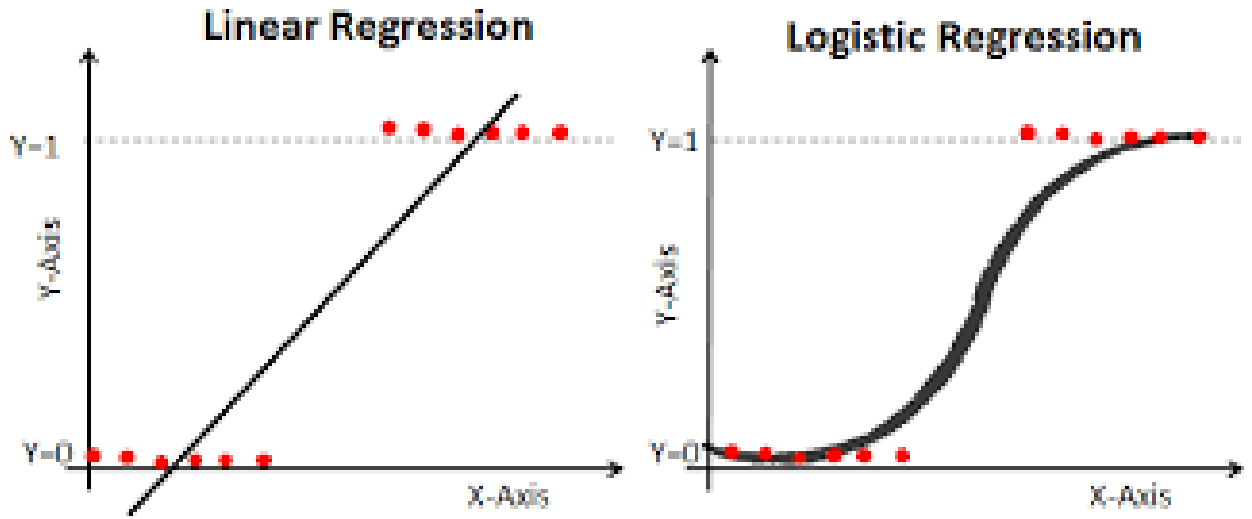
$$P(y_j/x_1, \dots, x_n) = \beta P(y_j) \prod_{i=1}^{n} P(x_i/y_j)$$

- y_j is a learning class (troll, non-troll)
- x_i are nominal attributes of troll's posting
- Naive Bayes is often applied as a baseline for text classification

Faculty of Electrical Engineering and Informatics

Logistic Regression

- Statistical technique to estimate parameters of a logistic model
- Linear combinations of independent variables are transformed using a specific type of logistic function - sigmoid function



- Computes the probability of a class - when the probability is greater than the given threshold, the example is classified to the class

Data Description

The service Export Comments was used to obtain data from Facebook

- the data are related to the SARS-Cov2 coronavirus pandemic in Slovakia
- comments were downloaded from the discussions on the Facebook profiles of the internet news portals “aktuality.sk” and “dennikn.sk” (portals with a significant impact)
- the comments were downloaded two days after the publication of the given paper (to ensure a certain consistency and authenticity of the data)
- we focused only on posts with the number of comments higher than eighty (since the probability of the occurrence of a troll increases with the number of reactions)
- data were downloaded for a period of one month since the outbreak of the pandemic in Slovakia

Data Description

The extracted text data were transformed to a training set with following attributes:

- Number of characters in the post
- Words count
- Average length of words
- Number of capital letters in the text
- Number of numbers in the text
- Number of “I like”

Another useful information was a measure of negativity, provocativeness, or hater in the comments of contributor:

- Positive comment
- Neutral comment
- Negative comment
- Provocative comment
- Strongly provocative
- Hateful comment.

Models Building

- The **Sci-kit package** was used for models training (SVM can work with several types of Kernel function.)

- Feature representation in the form of selected attributes
- Feature representation **enriched** by weights of words from comments of a contributor when **TF-IDF** weighting was applied

- **SVM, MNB and LR** models for recognition of troll authors were learned and **tested**

Models Testing

Results of testing machine learning models (SVM, MNB and LR) with and without the enrichment of data using TF-IDF weighting scheme

without TF-IDF		Precision	Recall	F1-rate
SVM	troll	**1,00**	0,08	0,15
	non-troll	0,56	**1,00**	0,72
MNB	troll	0,63	**0,92**	0,75
	non-troll	**0,89**	0,53	0,67
LR	troll	0,67	**0,89**	0,76
	non-troll	**0,90**	0,69	0,78

with TF-IDF		Precision	Recall	F1-rate
SVM	troll	0,36	0,38	0,37
	non-troll	0,43	0,40	0,41
MNB	troll	0,60	**0,92**	0,73
	non-troll	0,88	0,47	0,61
LR	troll	0,67	0,77	0,72
	non-troll	0,75	0,67	0,71

Conclusions

- **To identify a troll contributor**, 3 machine learning classifiers were tested
  - **SVM** model had the best results according to
    - Precision for troll recognition (FP)
    - Recall for non-troll recognition (FN)
  - **MNB** model on the contrary had the best results according to
    - Recall for troll recognition (FN)
    - Precision for non-troll recognition (FP)
  - Results of **LR** model was similar than MNB model but a little worse
- Enrichment of feature representation using TF-IDF was not beneficial
- For future, we would like to include some specific methods of machine learning with the potential to improve results
  - ensemble learning, deep neural networks learning and unsupervised learning

Faculty of Electrical Engineering
and Informatics

THANK YOU FOR YOUR ATTENTION