Compare Differences Method for Toxic comment classification

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], Global Customer Service[b,]*

[a]1600 John F Kennedy Boulevard, Philadelphia
[b]360 Park Avenue South, New York

Abstract

Toxic comment is one popular problem in the social network today. Every day, a billion of users posts their data on online network platform includes their images, videos, stories, comments.... Detect offensive content like toxic comment and restricts user post it is one essential task. Machine learning algorithm is one common approach, however Researcher gets tucked with many choices, it needs to decide the good approach for specific problem. To achieves the purpose above, I try many approach on the real dataset to present their effectiveness also the comparison based on quantity result also be listed in this research.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template
2010 MSC: 00-01, 99-00

1. Introduction

. Online shopping, social networking, chatting application is and on going to exploit until now. Entertainment content site like film, music store also influence many users today. Moreover, they can chat, comment with in the difference point of views with uncontrolled emotion, they make the hate intention together.

[⋆] Fully documented templates are available in the elsarticle package on CTAN.
[*] Corresponding author
Email address: `support@elsevier.com` (Global Customer Service)
URL: `www.elsevier.com` (Elsevier Inc)
[1] Since 1880.

. There are some dataset online collecting toxic comment on Twitter or drawled using tool on internet. I use the public dataset on by Surge AI to construct this research. This dataset is balanced for two class, so I can focus more on model evaluation and construction.

2. Relative work

. They are some works on toxic classification. Tobias Bornheim et .all apply GBERT or GELECTRA for detect toxic on Germany Language. Several competitions on Kaggle provide well labeling dataset drawled from social platform. For Example, The Toxic Comment Classification Challenges come with 35000 Dollars for the winner. Aken give dept analysis based on error for toxic comment besides common challenges such as out-of-vocabulary words, Long-Range Dependencies, Multi-word phrases.

3. Toxic comment challenges

. The performance of the classifier can be affect by unstructured adjective of comments. Comment also has characteristic of Spoken Language, it makes comment can have predefined meaning in small context make classifier hard to learn. The Dataset in this comment also has some special characters.

- Acronyms: an abbreviation consisting of the first letters of each word in the name of something, pronounced as a word. For Example: AIDS is an acronym for "Acquired Immune Deficiency Syndrome". FuK is an acronym for "Fuck". One Acronyms can contain toxic intention to make the whole sentence negative meaning.
- Special symbol: @hashtag or :)) (smile). It is too impossible if ignore all special impact to evaluate the hate score.
- Slang words: For example, the teen words can have same meaning with in other words.

https://github.com/tianqwang/Toxic-Comment-Classification-Challenge

3.1. Dataset Explore

. Dataset has been released on GitHub, it contains 1000 samples was labeled in 2 class: is not and not toxic. This is balancing dataset which has 50 percentage of each class.

4. Method and ensemble

4.1. Support Vector Machine

. Support Vector Machine attempt to minimize structural risk, attempt to limit on the generation error. In general, generation error come from overfitting on the dataset. It can be eliminated by constrain the decision boundary using a support vector.

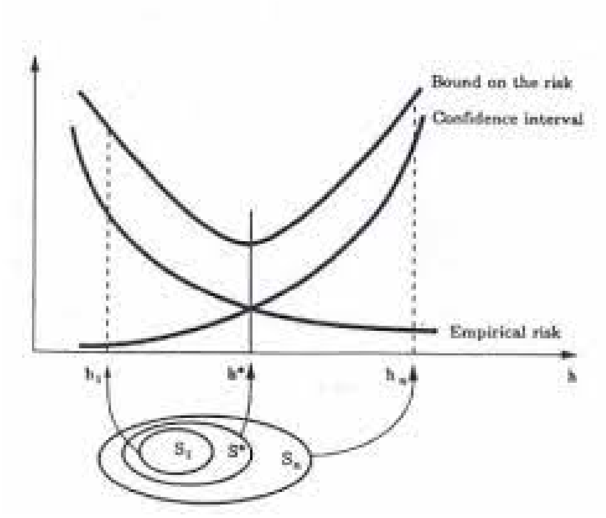

Figure 1: Structure Risk Minimization

4.2. Linear Regression

. Linear Regression is an algorithm used to modeling a linear relationship between a scalar and vector variable. This is simple approach compare deep learning approach and usually used for table dataset. Linear Regression make assumption that data point has linear relationship between its dimensions.

4.3. Naive Bayes

. Naive Bayes using statistic to make predict based on join distribution of data. Naive Bayes includes many sub-algorithm specify by the distribution function. For example, Multinomial Naive Bayes with Multinomial Distribution.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

4.4. Transformers

. Transformer is great deep model with introduction of Attention mechanism make simpler learning process than long short term memory. Transformer use multiples encoders and decoders to extract linguistic features from data.

4.5. Ensemble

. Ensemble method is a combination of many weak learners to make prediction. Naive Bayes, Bootstrap, Adaboost is can be seen as a kind of Ensemble. A tree is a combination of the rule extract from data.

. In general, weak learner is usually Decision Tree.

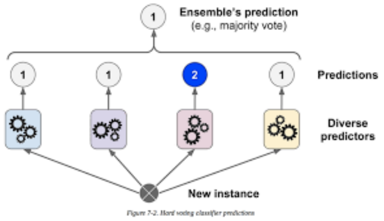

Figure 2: Ensemble

4.6. Rule-based classifier

. Contrast with learning base is rule based classifier. Simple rule base can fast and low memory consumption but complex to designed.

. Many researches use rule based to extract linguistic feature to feed into classifier, however automatically extract method like deep learning is more used approach.

5. Optimal Algorithm

https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/

5.1. Recall, F1-score, Precision, Accuracy

5.2. Time

5.3. Complexity

5.4. Interpretability

5.5. Space

6. Implementation

6.1. Preprocessing

. To achieve better result, the preprocessing stage is important, but the noisy in the dataset is make by people when labelling. Context also has limited influence because comment is not too long.

. I used NLTK includes Punkt, stop words, WordNet for cleaning text data. One by one, the sample was slug into following steps.

(1) Remove Associate links.

(2) Remove all special characters, just retain the alphabet characters.

(3) Tokenizing and lemmatizing

(4) Remove Stop words

(5) to reconstruct sample using space between words.

6.2. Feature Extraction

. I used Bag-Of-Word Technique for feature extraction. Dataset us divided into 80/20 for train set and test set.

7. Experimental Result

Method	Precision	Recall	F1-score
SVM	0.73	0.73	0.74
Linear Regression	0.68	0.79	0.73
Naive Bayes	0.81	0.8	0.8
Transformers	0.5	0.52	0.51
Ensemble (Bagging)	0.84	0.82	0.83

8. Conclusion and future work

. Hate speech detection is extending problem of toxic comment. Dataset for hate speech is common and has many resources, I will take this problem for the new part.

References

[1] Risch, Julian, et al. Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments. Universität Klagenfurt, 2021, doi:10.48415/2021/FHW5-X128.

[2] van Aken, Betty, et al. "Challenges for Toxic Comment Classification: An in-Depth Error Analysis." Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), Association for Computational Linguistics, 2018.