Toxic Comment Classification - Project Proposal

Shaishav Shah
CISE department, University of Florida
shah.sh@ufl.edu

***Abstract*—In today's world, social media has become one of the most efficient ways for people to communicate and discuss. People can share their ideas, their feelings, their opinions, the events happening in their life as well as react to the comments posted by other people. Considering all these advantages, social media may seem like a paragon for communication, but it's a double-edged sword as it can easily be abused by toxic comments from some users. While traditional moderators are in place to identify these types of comments and take appropriate actions, that is not the most efficient way for solving this problem.**

The goal of this project is to develop a machine learning[6] model that can autonomously identify and classify the toxic comments posted online. Through this project, I hope to gain a deeper understanding of the content introduced in the Pattern Recognition course.

I. Dataset

The dataset we would be using for this project is from one of the Kaggle competitions and can be found here: https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data. It contains a large number of Wikipedia comments which have been identified and labeled by human moderators for toxic behavior. The following six toxicity levels are defined: toxic, severe_toxic, obscene, threat, insult and identity_hate. As the comment here may belong to one or more toxicity levels, this is a multi-label classification type of problem. There are more than one hundred fifty thousand records of training and testing data available.

II. Plan of action

Following is the weekly plan in order to complete the project within the given deadline.

A. Week I - Pattern identification

- Study the dataset in order to identify the hidden patterns.
- Create visual representations (graphs) which will help in deducing observations and identifying important features to be used further in classification.
- Draft about the work done in the project report.
- Goal of this week is to have a good insight into the distribution of the dataset.

B. Week II - Preprocessing

[5]

- Perform Normalization on the dataset. Normalization generally refers to a series of related tasks meant to transform the data to a level where it is ready to be used by our model. The tasks include - 'Converting all text to uppercase or lowercase', 'Converting numbers into words', 'Removing punctuation and whitespaces', 'Removing stop words', 'Stemming', and 'Lemmatising'.
- Draft about the work done in the project report.
- Goal of this week is to have preprocessed data ready to be used by our model for training.

C. Week III - Literature review & Research

- Do literature review of various classification algorithms such as Binary relevance[9], Label power set[1], multi-label k-nearest neighbors[10], decision trees[7], and back-propagation multi-label neural networks[11].
- Study about various classifiers such as Multinomial Naive Bayes[4], Gaussian Naive Bayes[3] and Support Vector classification[8].
- Research about the python tools and/or libraries that can be used to solve this problem.
- Research about the tools for performance testing which will help in comparing these algorithms and choosing one over the others.
- Draft about the work done in the project report.
- By the end of this week, the development environment with all the tools necessary for model development will be installed and ready to be used.

D. Week IV - Model implementation

- Develop and train a model with a pair of algorithm and a classifier studied in the previous week. Repeat this for various combinations of algorithms and classifiers.
- Run the model with the given test data.
- Create a confusion matrix[2] and measure performances of various models and compare their accuracy.

E. Week V - Performance testing

- Compare various models considering their efficiency and accuracy and conclude which model performed best among the chosen ones.
- Perform final sanity check of the entire projectś implementation.
- Draft the project report in the IEEE format.

F. Week VI

- Complete final edit for the project report.
- Prepare PowerPoint presentation.
- Record and edit video demonstration, and upload it on YouTube.

Few days of this week are kept as a buffer in order to deal with any unplanned issues while performing the duties planned above.

REFERENCES

[1] Ziad Abdallah, Ali El-Zaart, and Mohamad Oueidat. An improvement of label power set method based on priority label transformation. 11:9079–9087, 01 2016.

[2] Tom Fawcett. Introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 06 2006.

[3] Hajer Kamel, Dhahir Abdulah, and Jamal Al-Tuwaijari. Cancer classification using gaussian naive bayes algorithm. pages 165–170, 06 2019.

[4] Ashraf Kibriya, E. Frank, Bernhard Pfahringer, and Geoffrey Holmes. Multinomial naive bayes for text categorization revisited. *Advances in Artificial Intelligence*, pages 488–499, 01 2004.

[5] Sotiris Kotsiantis, Dimitris Kanellopoulos, and P. Pintelas. Data preprocessing for supervised learning. *International Journal of Computer Science*, 1:111–117, 01 2006.

[6] Diego Oliva and Erik Cuevas. *An Introduction to Machine Learning*, pages 1–11. 11 2017.

[7] Lior Rokach and Oded Maimon. *Decision Trees*, volume 6, pages 165–192. 01 2005.

[8] Christian Ullrich. Support vector classification. 04 2009.

[9] Min-Ling Zhang, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. Binary relevance for multi-label learning: an overview. *Frontiers of Computer Science*, 12, 11 2017.

[10] Min-Ling Zhang and Zhi-Hua Zhou. A k-nearest neighbor based algorithm for multi-label classification. volume 2, pages 718 – 721 Vol. 2, 08 2005.

[11] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *Knowledge and Data Engineering, IEEE Transactions on*, 18:1338– 1351, 11 2006.