CE807 – Assignment 1 - Interim Practical Text Analytics and Report

Student ID: 2200940

Abstract

Text classification is a language processing and Text mining. A review based on Generic Text Classification from handcraft-rule to neural networks and spotting the relevant fields on Text Classification which includes Offensive Language, hate speech on social media were discussed. Some advantages and disadvantages of text classification were given.

1 Review of Generic Text Classification Methods (Task 1)

In the early days, text classification was done manually by experienced domain experts. That manual categorizing method produced quality results but it was very time-consuming. At the same time, they faced difficulties with larger datasets. Once the web pages were developed, they used to have more documents with larger datasets. At that stage, Automatic Text classification, such as rule-based(handcraft rule) and machine-learning methods jumped into the text classification process(Ignatow and Mihalcea, 2016) and (Aas and Eikvil, 1999). Text classification techniques have a wide range of applications, including E-Mail spam detection, sentimental analysis/opinion mining, gender classification, deception detection, and others. In order to reduce the burden of customer service, some researchers proposed an e-mail response template (Weng and Liu, 2004). Microblogging services, such as Twitter, Facebook, news, events, and private messages, contain significant amounts of raw emotional data that affect users. To address this issue, researchers proposed using a small set of domain-specific features extracted from the author's profile as well as his interests in short text classification and Bag of Words(Sriram et al., 2010). Researcher's found that much of the research work on some days around supervised learning techniques such as classification trees, Navies Bayes, Support Vector Machines, Neural nets, and ensemble methods. Because of increasingly large document collections, the supervised learning classifier's performance has degraded. A new approach was proposed for that problem to provide specialized understanding at each level of the document hierarchy, called Hierarchical Deep Learning for Text Classification (HDLTex) which employs stacks of deep learning architectures (Kowsari et al., 2017). Researchers encountered difficulties in identifying appropriate techniques for text classification, which required a deeper understanding of machine learning methods for accurate results. The survey was conducted in order to determine the limitations of each technique among various text classification algorithms (Kowsari et al., 2019). For text classification, several research studies have used deep learning approaches such as Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Papers on Graph Neural and Graph Convolution for text classification tasks were reviewed for clarity. They outperformed their traditional and deep learning-based counterparts on the reported data sets, according to their evaluation results. As a result, more neural network-based text classification applications and research works are expected in the future (Malekzadeh et al., 2021).

1.1 Critical Discussion (Task 1):

Text classification is based on Natural Language Processing. As discussed above, methods for text classification were developed with the development of technologies. There is a dramatic development of Text classification methods as

Handcraft-rule

↓

Machine Learning Models

↓

Deep Learning Models

↓

Neural Networks Models

These methods are maximum based on binary classification. Unsupervised learning models were used if the data is unlabeled and Supervised learning models were used if the data is labeled. Semi-supervised learning is a type of supervised learning problem that uses unlabeled data to train a model.For Text Classification,

2 Review of Offensive Language Detection Methods (Task 2)

Since all the given papers are focused on offensive language and/or hate speech detection, all papers' concept/intention was similar,i.e to detect offensive language and/or hate speech but their approach methods were different. So an overview of offensive language is given here in short and their approach methods were given below in detail. Offensive, Abusive, Insulting, and hate speech languages are very harmful and affect people's mental stage. This type of language is based on discriminating against an individual's ethnic background, race, gender, disability, and so on. These languages are seen publicly on social media, Facebook, Twitter, websites, online forums, news, etc. Sometimes, these languages are making a person commit suicide or become a murderer. For these reasons, much research work was done on Offensive Language Detection Methods, and also still working on it better.

Identification of specific fields and exploration of advanced methods:

In (Warner and Hirschberg, 2012), authors worked on a *Support Vector Machine classifier* to detect Hate speech. Firstly, a template-based strategy (Yarowsky, 1994) is used to generate features from a corpus and expand the feature set. After, the produced features were fed into the SVM classifier, where each feature is the dimension in a feature vector. Finally, generated each type of feature template strategy, implies six classifiers for each majority and gold copra. Through this process, the feature template successfully modeled hate speech as a classification problem. In (Waseem and Hovy, 2016), authors used a *logistic regression classifier* with cross-validation to test the influence of various features on prediction performance and to quantify their expressiveness. Next, they performed a grid search over all possible feature set combinations to find suitable features. Finally, found that using a character n-gram-based approach provides a solid foundation. Henceforth, the text classification model is used for Hate Speech detection. In (Malmasi and Zampieri, 2017), authors applied *standard lexical features* and a *linear Supervised classifier* to establish a baseline for this task. Here, mentioning the used three groups of features: surface n-grams, word skip-grams, and Brown clusters. Finally, they found a character 4-gram model's accuracy is best for their considered datasets. Therefore, the text classification model is used for Hate Speech detection. In (Vidgen et al., 2020), authors presented a *human-and-model-in-the-loop process* for collecting and training hate speech detection models in online. Firstly, they started their approach with four rounds of data generation and model training, but in principle, this could be continued indefinitely. In early rounds, models had lower accuracy but from later rounds models trained on data achieved higher accuracy. Also for each round annotators provided more challenging content in order to trick them. Finally, they showed the performance of target models improves as measured by their accuracy on the test sets. Therefore, the models trained on these dynamically generated datasets are much better at hate speech detection. In (Caselli et al., 2020a), authors introduced *HateBERT, a retraining BERT model* for abusive language detection in social media in English. Firstly, the collected Reddit Abusive Language English dataset (RAL-E) was split into training and testing sets to retrain and test the performance of the English BERT, the base-uncased model by applying the *Masked Language Model (MLM)* objective. Next verified the usefulness of HateBERT for detecting abusive language phenomena among three English datasets: OffensEval 2019 (Zampieri et al., 2019b), AbusEval (Caselli et al., 2020b) and HatEval (Basile et al., 2019). Finally, it showed that in-all datasets, the introduced HateBERT, a retraining BERT model for abusive language detection performed well.

2.1 Critical Discussion (Task 2)

The process of hate speech/Offensive Language Detection methods is similar to generic text classifications because all the selected/proposed detection methods are around Cleaning, Pre-processing, N-fold Cross-Validation, Hold-out tests, Feature selection/Extraction, Accuracy, Precision, Recall, Confusion Matrix and N-gram approach.

Reason for Selected methods:

- Support Vector Machine Classifier - because it is a binary classification problem where it identifies whether a word is hate speech or not.

- Logistic Regression Classifier - because a prediction function in logistic regression returns the probability of the observation being positive, Yes, or True, otherwise it is negative.

- Linear Supervised Classifier - because it has a label/target(i.e, hate speech or not) in the training set to train the model and to detect yes or no.

- Masked Language Model with retraining BERT Model - because it is a bi-directional and trained model to predict the required word based on its context.

3 OLID Dataset Characterization (Task 3)

- Offensive Language Identification Dataset (OLID) was made and collected by Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar from Twitter with Application Programming Interface (API).

- The original authors of OLID make this dataset publicly available online with annotation of type and target of offensive language for further research works. By referring/citing their paper, anybody can use OLID.

- The *topic* of the data is "Offensive Language Identification Dataset (OLID)" which is a large collection of English tweets. A lot of interesting research directions were *generated* from our study based on the performance of the different machine-learning learning models on OLID, which was the first dataset to contain annotation of the type and target of offenses in social media. The *Quantity* of OLID is 14,100 annotated tweets which are divided into a training partition of 13,240 tweets and a testing partition of 860 tweets. OLID was *organized* with a proposed three-level hierarchical annotation schema, such are Offensive Language Detection, Categorization of Offensive Language, and Offensive Language Target Identification, which makes it a useful resource for various offensive language identification and characterization tasks.In this article, they presented a new OLID dataset with annotation of the type and target of offensive language.

- This study is definitely relevant to my task since I am explored about different abusive and offensive identification from aggression to cyber-bullying, hate speech, toxic comments, and offensive language. From this study, I am familiarized with different aspects of text classification, which will be helpful for future works.

- Now, the OLID dataset is available publicly at `https://paperswithcode.com/dataset/olid`

- This OLID data was produced and collected from Twitter, Social Media.

- OLID was produced for predicting the type and target of offensive posts in social media because previous works on identifying offensive messages weren't cover many things.

- This proposed data set is trustable because the OLID was examined and calculated the performance of machine learning models in predicting offensive and non-offensive words on Twitter. Before OLID was proposed, annotators were labeled each tweet at all three levels of the annotation scheme such as

  1. whether a message is offensive or not
  2. what is the type of the offensive message
  3. who is the target of the offensive message

- OLID was produced at International Workshop on Semantic Evaluation (SemEval-2019 Task 6).

- Many researchers have used OLID for research works but as for now, there is no evidence regarding the change in OLID. They have only mentioned future additions to the dataset in the research paper. Note: They produced another new dataset named SOLID. (Zampieri et al., 2019a)

4 Summary

Day by day, Offensive content increased in the Internet Community. There were many research works done based on this issue and proposed Classifiers for Offensive Language Detection, were trained and evaluated by predicting the labels for the held-out test set. Nowadays, Automatic classifications can detect offensive language before it is published. Research on safety and security in social media has grown substantially in the last decade.

Discussion of the state-of-the-art :

The state-of-the-art Offensive Language based Text Classification is reviewed as follows: The first step is the Data collection, then the collected dataset should be analyzed in detail for Cleaning and Pre-processing. Next, data should be labeled and that labeled dataset is constructed into vector features. This data representation splits into training and testing sets with N-fold cross-validation. The selected Classifier would be trained with a training set and tested with a testing set. Trained Classifiers were assessed based on accuracy, precision, recall, and F1 score. After this evaluation, the trained model starts to predict the new dataset. Most of the models used small datasets so it is recommended to analyze those models with huge datasets.

Text Dataset
↓
Converted into vector features
ie., in numerical categorization
↓
N-fold Cross-Validation
↓
Training a classifier with training set
↓
Testing classifier with a testing set
↓
Predicting the new dataset

FLOW CHART OF THE OFFENSIVE LANGUAGE BASED TEXT CLASSIFICATION PROCESS WITH STATE-OF-THE-ART ELEMENTS

Advantages and Disadvantages:

- Text Classification methods are simple and the Offensive Language Detection methods are very useful in social media.
- Real-life Example, suppose we are planning to post some messages on social media (Twitter, Facebook, Instagram). If we type an offensive word in a message because of unfamiliar language, we are stopped by that social media with a small message which tells that the message is offensive. Then we can correct our words in a message before posting publicly. The main advantage of automatic text classification on offensive language detection is that we are warned/get to know before the negative thing happens.
- Most of the proposed methods are limited to a specific language, English. So it is necessary to implement it in multi-languages.
- Real-Life Example, Suppose we are having an AI Robot in our home and we are familiar with voice instructions. If our throat is not clear or if our pronunciation is wrong, that AI Robot misunderstood everything and the given instructions will be horrible. Sometimes, AI robots dominate humans, the important disadvantage is that we are used to depending on technologies for each and every simple action, which changes our character on waiting for google suggestions for text/messages.

Lessons learned:

- The landscape of text classification was explored.
- History of Text Classification from Handcraft-rule to Neural Networks was identified.
- Different research areas of text classification with a focus on offensive language and/or hate speech detection were identified.
- Learned about available software packages for automatic text classification.

References

Kjersti Aas and Line Eikvil. 1999. Text categorisation: A survey. Technical report, Citeseer.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020a. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020b. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Gabe Ignatow and Rada Mihalcea. 2016. *Text mining: A guidebook for the social sciences*. Sage Publications.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. 2017. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 364–371. IEEE.

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text classification algorithms: A survey. *Information*, 10(4):150.

Masoud Malekzadeh, Parisa Hajibabaee, Maryam Heidari, Samira Zad, Ozlem Uzuner, and James H Jones. 2021. Review of graph neural network in text classification. In *2021 IEEE 12th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*, pages 0084–0091. IEEE.

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. *arXiv preprint arXiv:1712.06427*.

Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas. 2010. Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Sung-Shun Weng and Chih-Kai Liu. 2004. Using text classification and multiple concepts to answer e-mails. *Expert Systems with applications*, 26(4):529–543.

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. *arXiv preprint cmp-lg/9406034*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983*.