REVIEW

Of the dissertation abstract submitted by Aigerim Bastarbekkyzy Toktarova for the degree of Doctor of Philosophy (PhD) in the field of Information Systems (8D06115), titled "Hate Speech Detection Using Natural Language Processing and Machine Learning Techniques in Online User Content"

The spread of messaging applications has caused some problems. One of these problems is hate speech. Hate speech causes problems that result in death as well as humiliating people. The thesis prepared by Aigerim Bastarbekkyzy Toktarova is about identifying messages that contain hate speech.

In the first stage, the candidate started working by collecting messages from various media (Tik-tok, Instagram, Facebook, Linkedln, Snapchat, Twitter) and researched methods that could reveal messages containing hate speech. The primary goal of the research is to develop a deep neural network model capable of automatically detecting hate speech in textual data.

The thesis consists of five chapters after Introduction as follows:

Introduction

1. Current Status and Issues of Detecting Prosecutive Words in Online Content
2. Literature Review of Scientific Research Work on Word Detection Algorithms
3. Using Natural Language Processing and Machine Learning Methods to Detect Words in Online Content
4. Research and Development of a Semantic Model for Detecting Words in Web Content
5. Comparative Results of The Research

In the first chapter of the dissertation, titled "Current Status and Issues of Detecting Prosecutive Words in Online Content" introduced key issues in identifying profanity and its classification, impact and types of profanity. A classified hate word dictionary was prepared. The classification of the hate words is:

- Word sequences (taboo) and obscene words as taboo words
- Terminology related to sexual orientation and gender discrimination
- Phrases containing words that are intended to cause harm and death
- A series of words with a discriminatory and insulting intent
- Words or phrases that combine concepts related to national discrimination and race
- Degrading people by comparing them to animals
- Phrases intended to discriminate against and mock people with physical and mental disabilities
- Extremism

In this chapter the impact and types of Kazakh swear words are also given.

The second section is devoted to the introduction of close and similar research. As a first step, the use of machine learning algorithms such as Naive Bayes, Support Vector Machine, Decision Machine, Random Forest, KNN, Logistic Regression are presented and compared as far as their performances. In the second step the concept of Neural Network classification and input parameter tuning are discussed. The methods used by the examined references and their achievements are given in summary. In the third step performance evaluation metrics for identifying words in a spoken language is discussed.

The third chapter is dedicated on to detect words in messages by using of NLP and ML methods. The role and importance of natural language processing and machine learning methods was emphasized. Then the fallowing methods were introduced.

- Stemming and lemmatization.
- Weighted measures of words
- K-means algorithm.

After that some application of machine learning methods (Decision Tree, Random Forest and Naïve Bayes, Logistic regression and K nearest neighbors) to identify words in a foreign language were introduced. Some text

classification methods such as Bag-of-Words, Word2Vec, Continuous Bag-of-Words, Continuous Skip-gram are introduced.

The fourth chapter is dedicated on research and development of a semantic model for detecting words in web content. Data are collected from Tik-tok, Instagram, Facebook, Tengrinews, Youtube. The following module were designed:

- Data collection modüle
- keyword search modüle
- document analysis modüle
- Building a parser to collect data

The following Works have been done for creating a semantic model:

1) Writing Kazakh letters by replacing them with Cyrillic letters
2) Using Latin letters when writing Kazakh comments
3) Using similes or metaphors to humiliate, harm personal dignity
4) Leaving comments using derogatory language words related to the region of residence
5) Writing without correcting spelling errors or using letters replace with a symbol
6) Create a Kazakh word by adding a suffix to a word from a Russian dictionary.

The description of the experiments which were done are:

1) Numbers and special characters are removed, only words are selected,
2) Suffixes and subordinate clauses are removed from the end of the word;
3) The presence or absence of each suffix or accepted word in the database is checked;
4) If several words are found in the database, the longest word is returned;
5) If there is no match in the database using the above methods, the search in the database starts from the beginning of the given word,
6) If it is not found in the database, the entered word itself is returned.

The fifth chapter is dedicated for comparative results of the research. Results of the proposed model: AUC ROC (Receiver Operating Characteristic – Area Under the Curve) are given. In the research, various machine and deep learning methods were tested. However, a hybrid model specially developed by the researcher showed the best results. This model combines the architecture of the modern contextual language model BERT with the ability to give more weight to important parts of the text using an additional attention mechanism integrated into it.

The proposed hybrid model outperformed traditional machine learning models by 17% for Kazakh, 15% for Russian, and 5% for English. These differences reflect the model's linguistic capabilities and context-awareness, demonstrating superior performance in detecting slurred speech.

Overall, this hybrid architecture based on BERT and the attention mechanism is considered one of the promising directions for achieving high performance in working with multilingual texts. It can find wide application in systems for automatic detection of slurred speech in online content and other areas of text analysis.

In conclusion, Aigerim Bastarbekkyzy Toktarova's dissertation, entitled "**Hate Speech Detection Using Natural Language Processing and Machine Learning Techniques in Online User Content**" submitted for the PhD degree in the field of Information System (8D06115), fully complies with the requirements of The State General Education Standard of the Ministry of Education of the Republic of Kazakhstan. I deem the candidate qualified for the conferral of the PhD degree in Philology.

Prof. Dr. Eşref ADALI, Istanbul Technical Univ.
21.04.2025