Project Proposal: Using Word Embeddings for Categorization and Analysis of Keyword Combinations Censored on WeChat

Nicola Lawford and Savanna Blade

Cyber Ethics and Digital Rights Portfolio
Engineers Without Borders U of T

July 1st, 2020

1 Introduction

Currently, the Citizen Lab's WeChat censorship monitoring system often finds 50+ new censored keyword combinations per day. The Citizen Lab also has a database of 180,000 blacklisted keywords from mobile games. These censored terms are manually classified into semantic categories for quantitative analysis of the most sensitive subject areas and the approach taken by China's information controls system in shaping how they are discussed online. However, manual categorization can introduce inconsistencies between researchers, and is labour-intensive; only 7,000 of the 180,000 mobile game keywords have been classified [1] A common technique in natural language processing is to represent words and phrases as semantically meaningful vectors or "embeddings"—thus, they can be added, subtracted, clustered, and semantically analyzed in various other useful ways. This proposal suggests a word-embedding methodology to enable automatic semantic categorization and further analysis of blocked keyword combinations.

Cyber Ethics and Digital Rights is a new portfolio at Engineers Without Borders U of T aiming to challenge the notion that engineering work and tech development are apolitical disciplines by engaging students with projects, skill development, and awareness about issues at the intersection of technology, social justice, and human rights. We propose undertaking this project in the 2020-21 academic year with an interdisciplinary team of 4-6 students.

2 Research Questions

The principal aim of this project is to develop word embeddings and an effective model architecture for automating the classification of censored keyword combinations into semantic categories. Secondary research topics include searching for meaningful new categories using unsupervised clustering, determining the distributions of sentiments and parts of speech in keyword combinations from each category, and documenting changes in these distributions over time. Throughout the project, we aim to use principles of design for edge cases in order to maximize performance on slang and code words used in attempt to subvert censorship.

3 Proposed Methods

3.1 Word Embedding

3.1.1 Why Represent Words as Vectors?

The simplest text categorization and sentiment analysis models are *lexicon-based*; that is, lists or lexicons of words associated with each category are compiled, and their relative occurrence in a text snippet determines the assigned category (often using a weighted sum or other methods) [2]. A lexicon for the category of keywords relating to Xi Jinping could contain "习" ("Xi"), "刁" ("Diao," a homograph for "Xi"), "包子" or "维尼" ("Buns" and "Pooh," slang nicknames poking fun at Xi Jinping). However, these models cannot classify snippets that do not contain any lexicon words, and miss latent sentiments and references to topics because they do not take context into account [3].

Word embedding methods use contextual occurrence data from a large corpus (containing millions of words) to represent words as semantically meaningful vectors. These vectors, usually having several hundred uninterpretable dimensions, can be added and subtracted to create new meanings: for example, the vector obtained from the operation $Paris - France + Italy$ is most similar to the vector representation of *Rome*. This allows for the cosine distance between two vectors to be used as a measure of similarity between words or phrases [4]. It also surpasses the limitations of lexicon-based methods by taking context into account.

3.1.2 Word Embedding Models

There are many methods for creating vector representations of words from a corpus. Latent Semantic Analysis (LSA) uses a "co-occurrence matrix" of words that appear together in documents, and singular value decomposition is used to reduce dimensionality [5]. Word2Vec, a model developed at Google, improves on the addition and subtraction properties of vectors from LSA. In this method, a neural network learns a projection of the word (one-hot encoded in the corpus vocabulary) onto a set number of dimensions, optimized to predict either a missing word from its context (Continuous Bag-of-Words) or the context words around a given word (skip-gram) [4]. Another method, Positive Pointwise Mutual Inference (PPMI) produces vectors with dimensionality equal to the size of the corpus vocabulary, with each word represented as a vector of its co-occurrence frequencies with all other words [6].

3.1.3 Chinese Word Embeddings

There are many existing pretrained Chinese word embeddings that could be effective for this task, including one developed by Tencent's AI Lab [7]. A comprehensive set of embeddings by Li et al. [8] is available on GitHub. They employed skip-gram and PPMI methods to a variety of corpora, including Weibo, various news sites, and Baidu Encyclopedia.

We hope to experiment with several word embeddings (trained with various models and corpora) for analysis of censored keywords. A relevant corpus to try is the set of articles tested on the Citizen Lab's WeChat censorship monitoring; thus, we are requesting access to this database in order to train our own embeddings.

3.2 Analysis Methods

3.2.1 Categorization Models

A simple categorization model applicable to word vectors is the k-Nearest-Neighbour model, which takes the k categorized vectors with the smallest distance from an uncategorized vector, and holds a vote to determine the categorization of the new vector [9]. However, a variety of more sophisticated text classification methods exist and are worth trying; for example, a combination of semantic clustering

with a convolutional neural network (CNN) was shown to perform well on short text snippets such as keyword combinations [10]. A benefit of using neural networks is that they output a classification confidence score in the form of a probability; low probability scores could be flagged for manual review.

Based on the size of existing datasets, we believe that automated categorization is achievable; currently, there are 2,174 COVID-19 related combinations blocked on WeChat in 6 parent categories, and datasets with similar sizes and numbers of categories have trained successful models. The aforementioned clustering + CNN model achieved accuracies of 85.5% on the Google Snippets dataset of 10,600 training snippets in 8 categories, and 96.8% on the TREC dataset containing 5,452 training questions in 6 categories. A combined lexicon and embedding model has achieved 78% on the MOBILE-SEN dataset of 2,315 Greek terms in 6 categories [3]. The 7,000 keywords in the database of mobile game blacklists that were manually placed in 6 categories should be sufficient to train a model for analysis of the remaining 170,000+ keywords.

For further exploration, unsupervised methods can be used to discover innate clusters in vectorized datasets. A common example is k-means clustering, which randomly initializes k clusters at single datapoints, and iteratively adds each data point to the cluster with the nearest mean [11]. We hope to test unsupervised clustering methods on the set of keyword combinations to see if meaningful categories naturally emerge.

3.2.2 Sentiment Analysis and Parts of Speech Tagging

Word embeddings allow for latent sentiment analysis, which can classify phrases as positive, negative, or neutral in tone, or rank them on a scale. Lin et al. trained several models for this task using a Word2Vec embeddings from a Chinese hotel review corpus, and found a support vector machine to be the most effective model architecture [12].

Parts of Speech (POS) tagging is a relatively popular research area in Chinese natural language processing, and many effective models exist. A model by Wang et al. uses a conditional random field model, using features from context words, a lexicon, and the results of unsupervised clustering [13].

Applying existing sentiment analysis and POS tagging models to the keyword data would give insight into whether censored keywords from each category were positive, negative, descriptive, or verb-based, and could track changes in the tone of censored discourse over time.

3.3 Designing for Edge Cases

In line with our aim to engage students with ethical issues in technical development, we hope to work with our teams to learn about ethical artificial intelligence. This involves designing machine learning models for "edge-cases:" data points for which data is limited or skewed, or that deviate from the general trend. In this case, these "edge-cases" are slang terms such as "刁," "包子" and "维尼," code words for Xi Jinping that are semantically unrelated and can be used in attempt to subvert censorship. In designing for edge-cases, we could borrow from design principles used in safety (e.g. autonomous vehicles) such as risk-sensitive performance criteria and sample weighting [14]. We could also use established slang detection methods such as lexicons and sounds-alike databases [15]; these would be useful for keyword combinations including components such as "鹿死," a homonym for "June 4" that literally translates to "Deer dead."

4 Data Access and Confidentiality

To make this project possible, we are requesting access to the database of keywords and combinations blocked on WeChat and mobile games, categorizations, and relevant researchers' notes that exist for each keyword or combination, as well as the database of articles that have been used as testing input during their collection. The data will not be shared beyond the authors of this proposal and the 4-6 project participants, and communication will be limited to encrypted platforms.

References

[1] Jeffrey Knockel, Lotus Ruan, and Masashi Crete-Nishihata. Measuring decentralization of chinese keyword censorship via mobile games. In *7th {USENIX} Workshop on Free and Open Communications on the Internet ({FOCI} 17)*, 2017.

[2] Jérémie Clos, Nirmalie Wiratunga, and Stewart Massie. Towards explainable text classification by jointly learning lexicon and modifier terms. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 19, 2017.

[3] Maria Giatsoglou, Manolis G Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos Ch Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings. *Expert Systems with Applications*, 69:214–224, 2017.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[6] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the eighteenth conference on computational natural language learning*, pages 171–180, 2014.

[7] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, 2018.

[8] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018.

[9] Naomi S Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[10] Peng Wang, Jiaming Xu, Bo Xu, Chenglin Liu, Heng Zhang, Fangyuan Wang, and Hongwei Hao. Semantic clustering and convolutional neural network for short text categorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 352–357, 2015.

[11] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[12] Yiou Lin, Hang Lei, Jia Wu, and Xiaoyu Li. An empirical study on sentiment classification of chinese review using word embedding. *arXiv preprint arXiv:1511.01665*, 2015.

[13] Yiou Wang, Yoshimasa Tsuruoka, Wenliang Chen, Yujie Zhang, Kentaro Torisawa, et al. Improving chinese word segmentation and pos tagging with semi-supervised methods using large auto-analyzed data. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 309–317, 2011.

[14] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[15] Alok Ranjan Pal and Diganta Saha. Detection of slang words in e-data using semi-supervised learning. *arXiv preprint arXiv:1702.04241*, 2015.