Chapter 5

Bidirectional GRU-Based Attention Model for Kid-Specific URL Classification

Rajalakshmi R.
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India

Hans Tiwari
School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

Jay Patel
School of Electronics Engineering, Vellore Institute of Technology, Chennai, India

Rameshkannan R.
School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India

Karthik R.
Vellore Institute of Technology, Chennai, India

ABSTRACT

The Gen Z kids highly rely on internet for various purposes like entertainment, sports, and school projects. There is a demand for parental control systems to monitor the children during their surfing time. Current web page classification approaches are not effective as handcrafted features are extracted from the web content and machine learning techniques are used that need domain knowledge. Hence, a deep learning approach is proposed to perform URL-based web page classification. As the URL is a short text, the model should learn to understand where the important information is present in the URL. The proposed system integrates the strength of attention mechanism with recurrent convolutional neural network for effective learning of context-aware URL features. This enhanced architecture improves the design of kids-relevant URL classification. By conducting various experiments on the benchmark collection Open Directory Project, it is shown that an accuracy of 0.8251 was achieved.

DOI: 10.4018/978-1-7998-1192-3.ch005
INTRODUCTION

Nowadays, everyone relies on the web and the internet has become the primary source for seeking information for various needs ranging from education to entertainment. Especially, children prefer to use the search engine for doing their school projects and assignments. In 2012, Yvonne et. al presented their study on the information seeking behavior of young children on the search engines. To enhance the searching for learning task, Ion Madrazo Azpiazu developed a web search environment that is suitable for children and teachers and analyzed the issues faced by them (Azpiazu et. al, 2017). To understand the search behavior of children, Nevena Dragovic analyzed the query logs and proposed a method to differentiate the general queries from the children queries (Dragovic, Azpiazu, and Pera, 2010). It is evident from the above listed research works that, the Gen Z kids rely on the web for seeking information and the role of web in their daily life. However, most of the children are not monitored during their surfing time and it is important to analyse the suitability of the web page. So, we are in need of an automated system that can help to determine the kids-relevant web page. The traditional method of content based classification is not suitable, as the kids related content of the web page is highly dynamic in nature. Many approaches have been explored in the literature to determine the suitability of the web pages for children (Patel and Singh, 2016 & Gyllstrom et. al, 2010). When the web page has animated videos or if the textual content is not available in such web pages, the traditional content based methods cannot be applied for classifying the web pages. Tatiana et al. (2010) have performed a survey on information retrieval system for children. They outlined the search query, search strategy and the navigation style used by children and discussed the need for a separate page ranking algorithm for children web pages. To rank the child related web pages according to the relevance, Gyllstrom et al. (2010) suggested a link based algorithm. Zhang et al. (2006) analysed the role of URLs in finding the objectionable content, but in their approach, they have combined the content and the URLs for classifying the web page. But an effective way to find the Kids-relevant web page is to make use of the links that they visit during their browsing time. Hence, a simple URL based classifier method is suggested to address the problems found in content based classification approaches.

In this research, we have proposed a deep learning method to classify Kids-specific web page by extracting features from URLs alone. For URL classification, tokens based features are suggested in literature (Kan, 2004 & Baykan et. al, 2011). The advantage of token based approach is that, high precision can be achieved as they reflect the topic of a web page. However, in many cases, the tokens may be misspelled or shortened for making the URL fancy / funny / catchy. So, we have integrated an auto spelling correction method in this system. The URL is a sequence of tokens that are delimited by some characters like: /, , , - etc. Also it may contain terms that are clueless to guess the topic it talks about. In some cases, the terms may be ambiguous, and may have different meaning based on the context. In traditional methods of URL classification, only the terms present in the URL are taken into account in the form of token features or n-gram features and the context of the terms are not considered. In the proposed work, in order to obtain context of the token, we have used word embedding. The URL contains a sequence of tokens and hence we applied a Recurrent Neural Network and tried to make use of sequence information also for classification of web pages. For example, consider an URL, https://www.education.com/kids-know-it/. The tokens ‘education’ and ‘kids’ are not adjacent to each to other, but if we use a Bi-directional Recurrent Neural Network, this information can be better utilized for classification. The earlier systems are limited by the hand-crafted features. In this research work, an attempt is made to