Chapter 6
Text Classification and Topic Modeling for Online Discussion Forums:
An Empirical Study From the Systems Modeling Community

Xin Zhao
University of Alabama, USA

Zhe Jiang
University of Alabama, USA

Jeff Gray
University of Alabama, USA

ABSTRACT

Online discussion forums play an important role in building and sharing domain knowledge. An extensive amount of information can be found in online forums, covering every aspect of life and professional discourse. This chapter introduces the application of supervised and unsupervised machine learning techniques to analyze forum questions. This chapter starts with supervised machine learning techniques to classify forum posts into pre-defined topic categories. As a supporting technique, web scraping is also discussed to gather data from an online forum. After this, this chapter introduces unsupervised learning techniques to identify latent topics in documents. The combination of supervised and unsupervised machine learning approaches offers us deeper insights of the data obtained from online forums. This chapter demonstrates these techniques through a case study on a very large online discussion forum called LabVIEW from the systems modeling community. In the end, the authors list future trends in applying machine learning to understand the expertise captured in online expert communities.

DOI: 10.4018/978-1-5225-9373-7.ch006

1. INTRODUCTION AND BACKGROUND

Systems modeling is the process of developing abstract models that represent multiple perspectives (e.g., structural, behavioral) of a system. Such models also provide a popular way to explore, update, and communicate system aspects to stakeholders, while significantly reducing or eliminating dependence on traditional text documents. There are several popular systems modeling tools, such as Simulink (MathWorks, 2019) and LabVIEW (National Instruments, 2019).

Laboratory Virtual Instrument Engineering Workbench (LabVIEW) is a system-design platform and development environment for a visual programming language from National Instruments. LabVIEW offers a graphical programming approach that helps users visualize every aspect of the system, including hardware configuration, measurement data, and debugging. The visualization makes it simple to integrate measurement hardware from any vendor, represent complex logic on the diagram, develop data analysis algorithms, and design custom engineering user interfaces. LabVIEW is widely used in both academia (Ertugrul, 2000, 2002) and industry, such as Subaru Motor (Morita, 2018) and Bell Helicopter (Blake, 2015). There are more than 35,000 LabVIEW customers worldwide (Falcon, 2017).

Text summarization refers to the technique of extracting information from a large corpus of data and represents a common application area of machine learning and natural language processing. With the increasing production and consumption of date in all aspects of our lives, text summarization helps to reduce the time to digest and analyze information by extracting the most valuable and pertinent information from a very large dataset.

There are two main types of text summarization: extractive text summarization and abstractive text summarization. Extractive text summarization is a technique that pulls keywords or key phrases from a source document to infer the key points from original documents. Abstractive text summarization refers to the creation of a new document for summarizing the original document. The result of abstractive text summarization may include new words or phrases not in the original documents.

To understand the current best practices and tool-feature needs of the LabVIEW community, we collected user posts from the LabVIEW online discussion forum. An online discussion forum is a website where various individuals from different backgrounds can discuss common topics of interest in the form of posted messages. Online discussion forums are useful resources for sharing domain knowledge. The discussion forums can be used for many purposes, such as sharing challenges and ideas, promoting the development of community, and giving/receiving support from peers and experts. Several researchers have identified benefits of online discussion forums from different aspects, such as education (Jorczak, 2014), individual and society development (Pendry & Salvatore, 2015) and socialization (Akcaoglu & Lee,