UMEÅ UNIVERSITET

Deep Learning for News Topic Identification in Limited Supervision and Unsupervised Settings

Arezoo Hatefi

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Salens namn eller beteckning, byggnad MIT.A.121, fredagen den 16 April 2024, kl. 13:15.
Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Eric Gaussier, Interdisciplinary Institute in Artificial Intelligence, University Grenoble Alps, Frankrike

Department of Computing Science

Organization
Umeå University
Dept. of Computing Science

Document type
Doctoral thesis

Date of publication
26 March 2024

Author
Arezoo Hatefi

Title
Deep Learning for News Topic Identification in Limited Supervision and Unsupervised Settings.

Abstract
In today's world, following news is crucial for decision-making and staying informed. With the growing volume of daily news, automated processing is essential for timely insights and in aiding individuals and corporations in navigating the complexities of the information society. Another use of automated processing is contextual advertising, which addresses privacy concerns associated with cookie-based advertising by placing ads solely based on web page content, without tracking users or their online behavior. Therefore, accurately determining and categorizing page content is crucial for effective ad placements. The news media, heavily reliant on advertising to sustain operations, represent a substantial market for contextual advertising strategies.

Inspired by these practical applications and the advancements in deep learning over the past decade, this thesis mainly focuses on using deep learning for categorizing news articles into topics of varying granularity. Considering the dynamic nature of these applications and the limited availability of relevant labeled datasets for training models, the thesis emphasizes developing methods that can be trained effectively using unlabeled or partially labeled data. It proposes semi-supervised text classification models for categorizing datasets into predefined coarse-grained topics, where only a few labeled examples exist for each topic, while the majority of the dataset remains unlabeled. Furthermore, to better explore coarse-grained topics within news archives and streams and overcome the limitations of predefined topics in text classification the thesis suggests deep clustering approaches that can be trained in unsupervised settings.

Moreover, to address the identification of fine-grained topics, the thesis introduces a novel story discovery model for monitoring event-based topics in multi-source news streams. Given that online news reporting often incorporates diverse modalities like text, images, video, and audio to convey information, the thesis finally initiates an investigation into the synergy between textual and visual elements in news article analysis. To achieve this objective, a text-image dataset was annotated, and a baseline was established for event-topic discovery in multimodal news streams. While primarily intended for news monitoring and contextual advertising, the proposed models can, more generally, be regarded as novel approaches in semi-supervised text classification, deep clustering, and news story discovery. Comparison with state-of-the-art baseline models demonstrates their effectiveness in addressing the respective objectives.

Keywords
Topic Identification, Data Clustering, News Stream Clustering, Semi-Supervised Learning, Unsupervised Learning, Event Topics, News Stories, Multimodal News, Document Classification, Document Clustering, Deep Learning, Deep Clustering, Pre-trained Language Models

Language
English

ISBN
print: 978-91-8070-342-0
PDF: 978-91-8070-343-7

ISSN
0348-0542

Number of pages
62 + 5 papers