ANNOTATED BIBLIOGRAPHIES

Frasconi, P., Soda, G., & Vullo, A. (2001). Text Categorization for Multi-page Documents: A Hybrid Naive Bayes HMM Approach. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 11–20). New York, NY, USA: ACM. https://doi.org/10.1145/379437.379440

In this article, Frasconi et al. present a method for classifying pages sequential OCR text documents into several assigned categories. The authors utilize a Naïve Bayes algorithm that factors in contextual information to improve accuracy during classification. The authors use data gathered from the University of Michigan digital library, Making of America to test their approach. The article is useful to my research topic because the authors the steps taking to build a text-based classifier. The main limitation of the article is that accuracy of the algorithm depends on the level of noise produced by the optical character reader(OCR); thus, the authors suggest the use of more sophisticated feature selection methods like the use of n-grams.

Grbovic, M., Halawi, G., Karnin, Z., & Maarek, Y. (2014). How Many Folders Do You Really Need?: Classifying Email into a Handful of Categories. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 869–878). New York, NY, USA: ACM. https://doi.org/10.1145/2661829.2662018

In this article, Grbovic et al. present an approach for automatically differentiating between personal and machine-generated emails as well as grouping messages into categories without human assistance. The authors' couple data gathered from a large-scale examination of user folder data and both online and offline learning classification components to determine user and machine-generated emails and create categories familiar to all users. The article is useful to my research topic because the authors describe the steps involved in building an email classification system. The main limitation of the paper is the classifier makes general email categorization thus the authors indicate that an investigation needs to be made in building sub-categories, e.g., banking under Finance.

Nenkova, A., & Bagga, A. (2003). Email Classification for Contact Centers. In *Proceedings of the 2003 ACM Symposium on Applied Computing* (pp. 789–792). New York, NY, USA: ACM. https://doi.org/10.1145/952532.952689

In this article, Nenkova et al. reviews an email routing and classification system that filters and classifies incoming emails upon their content for a contact center. The authors organize the system into two modules the first, a system that sorts and identifies emails that require immediate attention and the second, a system classifies messages. The authors use data from a discussion list from a web archive on the University of Washington to test the accuracy of the system by using a web archive from the. The article is useful to my research topic because the authors offer insight into the accuracy using Naïve Bayes and support vector machines approaches to building a classifier. The main limitation of the article is the difficulty in acquiring data due to privacy issue; thus, the authors spend a lot of time on trying to preprocess data they got.

Tam, T., Ferreira, A., & Lourenço, A. (2012). Automatic Foldering of Email Messages: A Combination Approach. In *Proceedings of the 34th European Conference on Advances in Information Retrieval* (pp. 232–243). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-28997-2\_20

In this paper, Ferreira et al. present an automatic email classification system based on supervised learning algorithms. The authors use a combination of the subject, body, date and other fields of an email to produce a classifier for automating email categorization. The article is useful to my research because the authors present a novel approach to analyzing emails to determine a category they fit in. Although the method yields significant improvement, the approached used three classifiers to process the subject, body, and participant of the message before collating in another classifier making the process intensive and time-consuming. Considering this, research has to be made into improving the algorithm while maintaining the accuracy.

T. Ayodele, R. Khusainov, & D. Ndzi. (2007). Email classification and summarization: A machine learning approach. In *2007 IET Conference on Wireless, Mobile and Sensor Networks (CCWMSN07)* (pp. 805–808). https://doi.org/10.1049/cp:20070271

In this paper, Ayodele et al. propose a system that reduces email overload by classifying emails based on user activities and summarizes emails to manage email information.

The authors achieve this by using an unsupervised learning algorithm to study email’s content and subject. The article is useful to my research because it uses an unsupervised learning approach to solve the problem of email classification. A limitation of this approach to email classification is the results produced are less accurate compared to the supervised learning approach usually used in classification. Considering this, the authors suggest an approach that further extracts essential words in the document to provide a better summary than running the unprocessed message in the classifier. This paper will not form the basis of my research. However, it provides useful information concerning text categorization.