CS60092 - Information Retrieval

Proposals for Term Project - Spring 2023

Group Members : Ashwani Kumar Kamal, Hardik Pravin Soni, Shiladitya De, Sourabh Soumyakanta Das

---

Project 1:

Title: Offensive query detection (on reddit/Twitter dataset) and generalization to multilingual settings.

Propositions:

1) Automatically classify the tweets into three categories: hateful, offensive and clean.
2) Test and compare various models for Hate-Speech detection on the basis of Precision, Recall and F1 score.
3) Show how such a model can be extended to generalize into multilingual settings without using multilingual models like XLM-R and mBERT.

Ideas to try out:

1) Existing model works on using Masked Language Model and training BERT on Reddit database with various parameters.
2) Ensembling various models of hate speech detection or offensive query detection to detect them.
3) Incorporating lemmatization and identifying various hateful-sentence structures of offensive queries and leveraging them.
   Eg: *I < intensity> <Hate - word> <target>* [2]
4) Extracting hateful posts or queries by comparing hateful and non-hateful communities.
5) Analyzing replies to various queries on Reddit and Twitter threads to identify hateful queries.
6) Comparing our model with existing models.

References:

1) https://arxiv.org/pdf/2010.12472.pdf
2) https://cse.iitkgp.ac.in/~mainack/publications/hatespeech-tham-2018.preprint.pdf
3) https://cse.iitkgp.ac.in/~mainack/publications/hatespeech-ht-2017.pdf

Project 2:

Title: Query-by-Example for Scientific Article Retrieval

Propositions:

1) Classification of documents into 4 categories, i.e., near identical, similar, related and unrelated using the background/objective, method or results as queries.
2) Using BERT and related models for finding similarities in given inputs and then performing classification on that basis.
3) Validating on the CSFCube annotated dataset.

Ideas to try out:

1) Using N-word stats which works on the basis of frequent bigrams or trigrams or even more appropriate would be wordset in a corpus.
2) Instead of searching for some specific keywords (scientific in this context), we can search for their related words (not just similar) as well, which can be implemented using YAGO, which would be more helpful in finding related articles.
3) We can also use citations/references to refer to even more scientific articles for proper context analysis. This can also help in making a better page ranking algorithm.
4) Collaborative filtering and applying a probabilistic model to this can also help cater similar queries with similar articles.

References:

1) https://arxiv.org/pdf/2103.12906.pdf
2) https://www.semantic-web-journal.net/system/files/swj169_1.pdf
3) Collaborative topic modeling for recommending scientific articles (acm.org)

Project 3:

Title: E-commerce Clothing Product Categorization with limited data.

Propositions:

1) Matching images with categories by training the models with limited amounts of data.
2) Multimodal search : Giving images to retrieve the serialized sequence of categories and vice-versa.
3) Modifying the existing model to allow it to train on a limited dataset.
4) Implement contrastive learning to enhance the model to work with less data.

Ideas to try out:

1) Use of PCA to reduce the computation overhead by reducing dimensions of the images.
2) Introducing the idea of cross entropy loss.
3) Incorporating siamese neural networks as they work on less datasets.
4) Incorporating semantic embeddings and use of triplet loss in siamese networks for image classification

References:

1) https://towardsdatascience.com/a-friendly-introduction-to-siamese-networks-85ab17522942
2) https://medium.com/product-categorization/product-categorization-introduction-d62bb92e8515
3) https://towardsdatascience.com/image-similarity-using-triplet-loss-3744c0f67973
4) https://aclanthology.org/C16-1051.pdf
5) https://arxiv.org/pdf/2202.02098.pdf