Exemplo gerado para uma seção intitulada “text representation for ranking” de um survey sobre “neural information retrieval”.

Cada seção está separada por linhas de pontos. O texto em cor preta foi o texto gerado. O texto em tons de bege foi o texto usado como base para o texto gerado.

.....
.....
.....

Title: Text-Image Matching for Cross-Modal Retrieval via Graph Neural Network

In recent years, there has been a growing interest in cross-modal retrieval, which aims to retrieve relevant information from one modality (e.g., text) given a query from another modality (e.g., image). This task is challenging due to the inherent semantic gap between different modalities. To bridge this gap, various approaches have been proposed, including the use of graph neural networks (GNNs) for text-image matching. In this section, we review the literature on text-image matching for cross-modal retrieval via GNNs.

One popular approach for text-image matching is to represent textual and visual data as embeddings in a shared latent space. These embeddings capture the semantic information of the data and enable effective matching between different modalities. One such method is Sentence-BERT [REF0], which utilizes Siamese BERT networks to generate sentence embeddings. By training the network on a large amount of sentence pairs, Sentence-BERT learns to encode the semantic similarity between sentences. This approach has been shown to achieve state-of-the-art performance in various natural language processing tasks.

Another approach is to leverage graph neural networks (GNNs) for text-image matching. GNNs are powerful models that can capture complex relationships between entities in a graph structure. In the context of cross-modal retrieval, GNNs can be used to model the relationships between words in text and visual features in images. For example, Yu et al. [REF2] proposed a neural pseudo-relevance feedback (PRF) approach that uses a feed-forward neural network model to aggregate the query and feedback document relevance scores and provide the target document's relevance score. This method utilizes graph transformers to capture the PRF signals from the initial retrieved results. Similarly, Wang et al. [REF2] proposed a clustering method to gather the relevance signals from PRF documents. These methods have shown remarkable improvements in cross-modal retrieval, but their efficiency is significantly affected, making them computationally infeasible for many practical applications.

In addition to GNNs, other models based on BERT have also been applied to text-image matching. BERT (Bidirectional Encoder Representations from Transformers) is a pre-training model that has achieved state-of-the-art performance in various natural language

processing tasks [REF6]. For example, dense retrievers [REF2] often utilize a BERT-based dual-encoder to encode queries and passages into a shared embedding space. These models have shown great effectiveness and efficiency in various tasks and datasets.

Overall, text-image matching for cross-modal retrieval via graph neural networks is an active area of research. Various approaches, including Sentence-BERT, GNNs, and BERT-based models, have been proposed to address the challenges of cross-modal retrieval. These models have shown promising results in capturing the semantic relationships between text and images. However, there is still room for improvement in terms of efficiency and scalability. Future research could focus on developing more efficient and scalable models for text-image matching in cross-modal retrieval tasks.

References:
[REF0] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT Networks. arXiv:1908.10084.
[REF2] Yu et al. (Year). Title. Conference/Journal.
[REF6] BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, pages 4171–4186.

References given to GPT:
[REF0] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: Nils Reimers and Iryna Gurevych. 2019. SentenceBERT: Sentence Embeddings using Siamese BERTNetworks. arXiv:1908.10084 [cs], August. arXiv: 1908.10084. Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. arXiv:2004.09813

[REF1] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: In the sequence, we also added an SVM classifier, to RBF kernel and C = 2 penalty to the ensemble, making Logistic Regression our meta-classifier. The training set was divided into 90% for training/validation and 10% for test set. Models were 4 https://scikit-learn.org/stable/ trained in the training/validation set using 10-fold cross-validation. (Han et al., 2011). 6 Results Tables 2 and 3 show the performance and settings of each classifier in the training/validation and test sets, respectively. During training, best results were observed without preprocessing, for RF and LR, whereas NB showed better results with preprocessing. These results, however, were very close to each other, ranging from F1=0.69 to F1=0.71.

[REF2] - paperID: 471dea6589d6f19e78db1f47fbc7cff0d9f1aab3
Title: Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback: A Reproducibility Study Chunk of text: proposed a neural PRF approach that uses a feed-forward neural network model to aggregate the query and feedback document relevance scores and provide the target document's relevance score. Yu et al. utilises graph transformers to capture the PRF signals from the initial retrieved results, and Wang et al. proposed a clustering method to gather the relevance signals from PRF

documents. These methods show remarkable improvements, but the efficiency is significantly affected, such as BERT-QE inference requires 11.01x more computations than BERT alone, making these models computationally infeasible for many practical applications. Recently, dense retrievers [29,16,7,8,6] have been attracting a lot of attention from researchers. These models, which often utilise a BERT-based dual-encoder to encode queries and passages into a shared embedding space, have shown great effectiveness and efficiency in various tasks and datasets.

[REF3] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf
Title: Learned Text Representation for Amharic Information Retrieval and Natural Language Processing Chunk of text: The performance of our models can be compared to some other languages' models using Tables 6 and 13. Table 13. Performance of word embeddings and BERT models on some tasks on some languages. Language Down Streaming Task Model Performance English Document classification BERT 0.96 (F1-score) Chinese Document classification BERT 0.97 (accuracy) English Ad hoc retrieval word2vec 0.48 (NDCG) English Query expansion word2vec GloVe fastText 0.086 (precision) 0.087

[REF4] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: Elijah Mayfield and Alan W Black. 2019. Stance classification, outcome prediction, and impact assessment: Nlp tasks for studying group decisionmaking. In Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 65–77. Stefan Ollinger, Lorik Dumani, Premtim Sahitaj, Ralph Bergmann, and Ralf Schenkel. 2020. Same side stance classification task: Facilitating argument stance classification by fine-tuning a bert model.

[REF5] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf
Title: Learned Text Representation for Amharic Information Retrieval and Natural Language Processing Chunk of text: . Some of these models were trained for cross-lingual purposes and are not usable for the needs of most NLP tasks. Moreover, most of them are not publicly accessible. Because of this, Amharic NLP tasks have been performed using classical text representations such as stems and roots [42,43], and the impact of learned text representations on roots, stems, and words to the development of various applications is not yet investigated. Thus, the construction of pre-trained Amharic models is a long sought resource for the research community. In view of this, the major contributions of this work are: (i) construction of pre-trained Amharic models and publicly sharing them to the research community; (ii) fine-tuning the pre-trained models for NLP and IR tasks; and (iii) investigation of the effects of roots, stems, and surface words on learned text representations.

[REF6] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL, pages 4171– 4186. Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel

Preot¸iuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 843–854, Berlin, Germany, August. Association for Computational Linguistics.

[REF7] - paperID: 7b8fe8c28a371120b4479540b2c8a0f7c5af25bf
Title: Learned Text Representation for Amharic Information Retrieval and Natural Language Processing Chunk of text: If a word is unseen during training, fastText segments a word into n-grams and generates its embedding. As a result, it helps to embed rare words, misspelled words, and words that do not exist in corpora but are found in the topic set. For example, the query term አገልግሎት /ʔəgəlɨgɨlotɨ 'services'/ is not found in the corpora, and thus the word2vec and GloVe models do not return any expanded terms. However, fastText returns the words አገልግሎ /ʔəgəlɨgɨlo/, አገልግሎትም /ʔəgəlɨgɨlotɨmɨ/, አገልግሎትን /ʔəgəlɨgɨlotɨnɨ/, አገልግሎትና /ʔəgəlɨgɨlotɨna/, አገለገሎት /ʔəgəlɨgəlotɨ/, አገልግሎቱንና /ʔəgəlɨgɨlotunɨna/, አገለገሎቱን /ʔəgəlɨgəlotunɨ/, አገልግሎቱን /ʔəgəlɨgɨlotunɨ/, and አገልግሎቱም /ʔəgəlɨgɨlotumɨ/, which are variants referring to the concept "serve". Furthermore, fastText based on skip-gram outperforms the baseline retrieval performance reported in . The effectiveness of the Amharic retrieval system without and with query expansion using fastText is presented in Table 12.

[REF8] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: The dataset hosted by SardiStance has tweets in Italian language about Sardines movement. The total tweets are about 3,242 instances out of which, training set has 2,132 and testing will have 1,110. The three stances are Against, Favor and Neutral about the Sardines movement with 1,028, 589, 515 instances respectively. 3.2 Model Construction The models are built in Python and used GPU system with NVIDIA GTX1080 for running the experiments. The features are extracted from the Italian tweets about Sardines movement to construct the model and the same is evaluated for performance using the tweets meant for testing. Feature engineering in our work includes both226 via the explicit features and also using a deep learning model that does the same.

[REF9] - paperID: 336e531a59cafbe215b950fd749bca866b89cea0
Title: SNK @ DANKMEMES: Leveraging Pretrained Embeddings for Multimodal Meme Detection (short paper) Chunk of text: The goal of the first subtask, Subtask A - Misogyny & Aggressive Behaviour Identification, is the identification of misogynous speech in tweets, and in case of misogyny, the classification of an aggressive language. Subtask B - Unbiased Misogyny Identification, aims at classifying misogynous speech while guaranteeing the fairness of the model (in terms of unintended bias) on a synthetic dataset. The unintended bias is a known phenomenon in natural lan1 https://www.theverge.com/2020/3/5/21166940/twitterhate-speech-ban-age-disability-disease-dehumanize, https://www.theverge.com/2020/8/11/21363890/facebookblackface-antisemitic-stereotypes-ban-misinformation, https://www.theguardian.com/technology/2020/jun/29/redditthe-donald-twitch-social-media-hate-speech49 guage models and recent

works address its identification and mitigation (Dixon et al. (2018), Nozza et al. (2019), Kennedy et al. (2020)). In this work, we describe our solution to address the AMI shared task. We propose a multiagent classification. The system uses recent Sentence Embedding techniques to encode tweets and a SVM classifier to produce initial labels.

.....
.....
.....

Title: Deep Learning Models for Text Representation in Ranking

Deep learning models have gained significant attention in the field of natural language processing (NLP) due to their ability to capture complex patterns and representations in textual data. In the context of ranking, text representation plays a crucial role in determining the relevance and importance of documents. This section explores the use of deep learning models for text representation in ranking tasks.

One approach to text representation in ranking is the aggregation of features along the temporal dimension [REF0]. This approach involves aggregating frame-level features, such as object, motion, face embeddings, by averaging them along the temporal dimension to produce a single feature vector per video. For other features like speech, audio, and OCR, the NetVLAD mechanism has been proven effective for the retrieval task [REF0]. This mechanism involves applying linear projections to transform the time-aggregated embeddings into the same dimensionality, enabling the aggregation of multiple features [REF0].

Graph-based models have also shown promise in text representation for ranking tasks. Hypergraphs, which are a generalization of graphs, have been explored for entity ranking [REF1]. Hypergraphs allow for the connection of an arbitrary number of nodes through hyperedges, which can be represented as tensors. Tensor factorization techniques have been applied to hypergraphs to extract meaningful representations for ranking tasks [REF1]. Additionally, the introduction of sentence, paragraph, or passage hyperedges has been proposed to improve the search process by avoiding unrelated directions [REF2].

In the domain of code ranking, deep learning models have been developed to determine the similarity between pieces of code [REF3]. The DeepCS approach utilizes joint vector representation to embed information from NL-query and code fragments, while the UNIF approach simplifies inputs and achieves better performance with a simpler model [REF3].

Evaluation of ranking models is crucial to assess their effectiveness. In entity ranking tasks, the effectiveness of different ranking alternatives has been explored, with findings suggesting that the number of documents citing an entity is more relevant than the frequency of entity citations [REF4]. The maximum entropy algorithm, combined with a scoring function that considers entity frequency and document frequency, has shown promising results [REF4].

Consolidating models is an important aspect of deep learning for text representation in ranking tasks. Unified models that consolidate different approaches, such as physics-based models and machine learning models, have been proposed [REF5]. These unified models aim to leverage the strengths of different techniques to improve ranking performance.

In recent years, deep learning solutions based on transformer networks have been used to enrich statistical information about terms and expand the collection of documents [REF6]. These approaches, such as DeepCT and doc2query, reweight terms occurring in documents to highlight important terms, resulting in augmented document representations that can be indexed using traditional inverted indexers [REF6].

The number of propagation layers in graph neural networks (GNNs) has been found to impact performance in ranking tasks [REF7]. Stacking multiple propagation layers allows for the capture of higher-order semantic relationships between query, item, and tag, leading to improved performance [REF7].

Probabilistic models, such as language models and divergence from randomness, have been employed to handle multi-keyword queries and consider query-independent evidence [REF8]. These models take into account the prior probability of a term and measure the information gain of a term given a document, providing a broader notion of relevance [REF8].

Entity-oriented search has also been explored for ranking tasks, addressing challenges in searching unstructured and structured data [REF9]. Semantic search approaches based on conceptual graphs and RDF graphs have been proposed, measuring the similarity between conceptual graphs based on the similarity between their nodes and edges [REF9].

In summary, deep learning models offer promising approaches for text representation in ranking tasks. Aggregating features, leveraging graph-based models, consolidating different techniques, and utilizing probabilistic models are some of the strategies employed to improve ranking performance. These models provide valuable insights into the representation of textual data for effective ranking.

References given to GPT:
[REF0] - paperID: cb89db971ca84b50facbfc0d4c6aa44f42894126
Title: Multi-Feature Graph Attention Network for Cross-Modal Video-Text Retrieval Chunk of text: Each element of this collection is then aggregated along its temporal dimension, producing a fixed-length embedding per video n I (1) , .., I (M) o . For temporal aggregation function, we adopt a simple approach to aggregate the features. For object, motion, face embeddings, we average the frame-level features along the temporal dimension to produce a single feature vector per video. For speech, audio, OCR features, we adopt the NetVLAD mechanism proposed by Arandjelovic , which has been proven effective for the retrieval task . Multi-Feature Graph Attention Module: Once the time aggregated embeddings

are obtained, we apply linear projections to transform these embeddings into the same dimensionality. These projected video feature embeddings can be written as: H = {h1, h2, ..., hM } , (1) where hi ∈ R F , and F is the number of features. To aggregate these multiple features, we first construct a multifeature graph for each video.

[REF1] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: The graph-based approach outperformed the vector space model for both the Chinese and English test collections, and it even outperformed the Google algorithm. 2.2.7 Hypergraph-based models Hypergraphs are a generalization of graphs, where edges (or hyperedges) can connect an arbitrary number of nodes — undirected hyperedges are represented by a set of nodes, while directed hyperedges are represented by a tuple of two sets of nodes. When all hyperedges in a hypergraph contain the same number k of nodes, the hypergraph is said to be k-uniform. In that case, it can be represented as a tensor of k dimensions, each of size |V|. In Section 2.2.5, we had covered tensor factorization over a tensor of entity relations for different predicates. Exploring analogous methods based on hypergraphs might also wield interesting results.

[REF2] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: The constraints provided by the hypergraph-of-entity are still not enough, in particular to support search using random walks over a collection of news articles. Several approaches might be taken to improve this, namely introducing sentence, paragraph or passage hyperedges in order to avoid taking steps into unrelated directions (such as “megan”). Obviously, de2239.3 universal ranking function evaluation Table 9.9: Best runs per team for TREC 2018 Common Core track. Team Run ID Type MAP UWaterlooMDS UWaterMDS_Rank Manual 0.4303 RMIT RMITUQVDBFNZDM1 Manual 0.3850 h2oloo h2oloo_enrm30.6 Automatic 0.3382 MRG_UWaterloo uwmrg Automatic 0.2761 Anserini anserini_qlax Automatic 0.2749 Sabir sab18coreE1 Feedback 0.2510 NOVASearch bt-BoWBoE Feedback 0.2468 UMass umass_sdm Automatic 0.2339 JARIR jarir_sg_re Automatic 0.2040 Webis webis-argument Automatic 0.1015 FEUP feup-run1 Automatic 0.0070 spite document scoring depending on r = 1,000 random walks for each seed node (frequently multiple entities for a single term), allowing such unrelated walks is still detrimental to the overall ranking. Furthermore, the hypergraph-of-entity does not support any type of document length normalization, which is also affecting the quality of random walks. We also did not use any stemming or lemmatization, since we wanted to leave room for the exploration of syntactic relations, which could only be extracted and modeled based on complete sentences.

[REF3] - paperID: 06227bc74bcee55471fb37bde0149b317f8a2014
Title: Enhancing Semantic Code Search With Deep Graph Matching Chunk of text: To determine how similar two pieces of code are to one another, Gu et al. offer the DeepCS approach, which uses joint vector representation to embed information of NL-query and code fragment. Cambronero et al. , proposed an approach named UNIF. This approach increases the efficacy of encoding and simplifies inputs. The UNIF use the concept of a bagof-words-based network that transforms docstring tokens and code snippets into

embedding vector using supervised learning. UNIF performs better and has a simpler model compared to DeepCS.

[REF4] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: In this section, we illustrate two evaluation approaches for entity ranking tasks. Komninos and Arampatzis presented a web application for entity ranking that receives a query in natural language and identifies the most relevant entities associated with the query. For evaluation, they used the topics from the entity ranking tracks from INEX 2009 and TREC 2010. They tested the effectiveness of eleven ranking alternatives, discovering that the number of documents that cite an entity is more relevant than the number of times the entity is cited in the documents. They also found that in the top-n retrieved documents, when considering a small n, document rank information has little influence over entity relevance. They verified that the best results were achieved when using the maximum entropy algorithm with a scoring function that combined the logarithmic entity frequency with the document frequency.

[REF5] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: 3 1.1.1 Information retrieval and the evolution of search . . . . . 4 1.1.2 The web, knowledge graphs, and real-world networks . . 7 1.1.3 Documents meet entities: the birth of entity-oriented search 9 1.2 The importance of consolidating models . . . . . 11 1.2.1 Unified models: from physics to machine learning . . . .

[REF6] - paperID: 5537feedc97256e81c6f1af66664dbcd19621d11
Title: ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval Chunk of text: Typically, these models identify and weight feedback terms that are frequent in the feedback documents and infrequent in the corpus, by exploiting statistical information about the occurrence of terms in the documents and in the whole collection. In all cases, the reformulated query is then re-executed on the traditional (so-called sparse) inverted index. Recently, deep learning solutions based on transformer networks have been used to enrich the statistical information about terms by rewriting or expanding the collection of documents. For instance, DeepCT reweights terms occurring in the documents according to a fine-tuned BERT model to highlight important terms. This results in augmented document representations, which can be indexed using a traditional inverted indexer. Similarly, doc2query and its more modern variant docT5query

[REF7] - paperID: 017386502557c27d4ffd575b17ed7c2aafed2d95
Title: Item Tagging for Information Retrieval: A Tripartite Graph Neural Network based Approach Chunk of text: Specially, we conduct experiments with the layer numbers in range of {1, 2, 3, 4}. Table 4 summarizes the experimental results, wherein TagGNN-X indicates the model with X layers. From the results, we have the following observations: • TagGNN-1 is obviously worse than TagGNN-2,3,4, indicating that only one propagation layer is not enough to reach an excellent performance. It is reasonable since one-layer GNN propagation can only capture the first-order neighbors' information. Hence, semantic relationships between query and query,

item and item, tag and tag are not explicitly used, resulting in unsatisfactory performance. So it is necessary to stack at least two propagation layers. •

[REF8] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: For multi-keyword queries, probabilities are multiplied. Language models take advantage of smoothing, usually Jelinek-Mercer or Dirichlet, in order to consider documents with missing query terms, or even documents with none of the query terms. A similar smoothing strategy is also explored in PageRank (see Equations A.2 and A.5 in Appendix A). By considering the prior probability of a term, we broaden the notion of relevance, taking into account query-independent evidence — i.e., a document might be relevant solely due to its terms, however it will often be less relevant than documents with a strong query-document relation. Divergence from randomness is a probabilistic model where we measure the information gain of a term given a document. This model is a generalization of Harter's indexing model , where two Poisson distributions (hence 2-Poisson) were combined in analogy to TF and IDF, using the notion of eliteness to describe documents with a more prominent presence of query terms, when compared to other documents. A similar approach is taken in divergence from randomness, but there is a wide range of models to select from, as opposed to only being able to use the Poisson distribution.

[REF9] - paperID: d121c33a5a0d8b6615d8581cfee8a941ebc7daed
Title: Graph-based entity-oriented search Chunk of text: Entityoriented search tackles this type of challenges, making search easier over unstructured and structured data. Zhu et al. and Zhong et al. have proposed an approach to semantic search for entity ranking, through the matching of a query graph and a resource graph. The idea was developed based on conceptual graphs , having a direct translation to RDF graphs1 . The conceptual graphs were built from natural language queries and documents via their prototype ALPHA . They measured the similarity between two conceptual graphs based on the similarity between their nodes and edges.

.....
.....
.....

Title: Textual Representations for Crosslingual Information Retrieval

In the field of Crosslingual Information Retrieval (CLIR), the representation of textual data plays a crucial role in achieving accurate and effective retrieval results. Various approaches have been proposed to represent text in a crosslingual setting, taking into account the challenges posed by different languages and the need for efficient and accurate retrieval. In this section, we discuss some of the key textual representations used in CLIR and their impact on retrieval performance.

One approach to textual representation in CLIR is the use of multilingual embeddings. Multilingual embeddings, such as

multilingual BERT (mBERT) and XLM-RoBERTa, have been widely adopted in CLIR tasks [REF8]. These models leverage pre-training over texts in multiple languages to learn crosslingual representations that capture semantic similarities across languages. By using these embeddings, CLIR systems can effectively bridge the language gap and retrieve relevant information across different languages.

Another approach to textual representation in CLIR is the use of neural network architectures, such as Transformers and RNNs. Transformers, in particular, have shown promising results in various natural language processing tasks, including machine translation [REF6]. These models can capture contextual information and dependencies between words, enabling them to generate rich representations for text. RNNs, on the other hand, are commonly used for encoding sequential data and have been applied to encode sentence embeddings in the same space as image representations [REF5]. These architectures provide a powerful means of representing text in a crosslingual setting.

In addition to neural network architectures, other methods have been proposed for textual representation in CLIR. For example, feature ensembles have been used to combine representations from different sources [REF2]. This approach aims to leverage the strengths of each source and improve the overall representation quality. Furthermore, weak supervision signals have been explored to approximate query-document relevance signals, such as pseudo relevance labels generated by unsupervised retrieval methods [REF9]. These signals, although noisy, can provide valuable information for training CLIR models and improving retrieval performance.

Evaluation of different textual representations in CLIR is typically done using metrics such as Normalized Discounted Cumulative Gain (NDCG) and Precision at K (P@K) [REF1]. These metrics measure the effectiveness of the retrieval system in ranking relevant documents higher than irrelevant ones. Experimental results have shown that certain textual representations, such as multilingual embeddings and neural network architectures, outperform traditional methods like Jaccard and Cosine similarity in terms of retrieval performance [REF1].

In conclusion, the choice of textual representation plays a crucial role in the success of Crosslingual Information Retrieval systems. Multilingual embeddings, neural network architectures, and feature ensembles are some of the approaches that have shown promising results in representing text in a crosslingual setting. By leveraging these representations, CLIR systems can bridge the language gap and retrieve relevant information across different languages effectively.

References given to GPT:
[REF0] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3
Title: UOBIT @ TAG-it: Exploring a Multi-faceted Representation for Profiling Age, Topic and Gender in Italian Texts
Chunk of text: For each social media platform, training and test data were provided. Furthermore, two cross-platform sub-tasks were introduced to test the systems’ ability to generalize across

platforms. The ultimate goal of HaSpeeDe 2 at EVALITA 2020 (Basile et al., 2020) is to take a step further in state-of-the-art HS detection for Italian. By doing this, we also intend to explore other side phenomena and see the extent to which they can be automatically distinguished from HS. We propose a single training set made of tweets, but two separate test sets within two different domains: tweets and news headlines. While social media are still one of the main channels used to spread hateful content online (Alkiviadou, 2019; Wodak, 2018), an important role in this respect is also played by traditional media, and newspapers in particular. Furthermore, we chose to include another HSrelated phenomenon, namely the presence of stereotypes referring to one of the targets identified within our dataset (i.e., muslims, Roma and immigrants).

[REF1] - paperID: 7715d2fc795a6406151b94924d9276939671f919
Title: TabSim: A Siamese Neural Network for Accurate Estimation of Table Similarity Chunk of text: Averaged over the three corpora, TabSim outperforms LR, RF, Cosine, Google Fusion and Jaccard in terms of NDCG@10 by 3.0% pp, 1.3% pp, 17.2% pp, 19.0% pp and 15.8% pp, respectively. TabSim also outperforms all competitors in terms of NDCG@5 by at least 4.5% pp, except RF. TABLE V 5F-CV NDCGS (%) FOR Jaccard, Cosine, Google Fusion, RF, LR AND TabSim OVER THREE CORPORA. BEST VALUE PER MEASURE IS IN BOLD. Corpora Method NDCG@5 NDCG@10 PMC Jaccard 93.10 94.66 Cosine 95.58 95.68 Google Fusion 94.51 95.04 RF 90.53 92.03 LR 92.11 93.13 TabSim 93.76 94.57 arXiv Jaccard 40.53 41.09 Cosine 35.03 36.18 Google Fusion 29.17 32.11 RF 81.07 82.26 LR 62.25 72.48 TabSim 74.15 82.71 Wikipedia Jaccard 91.38 91.45 Cosine 91.06 91.14 Google Fusion 90.13 90.28 RF 96.46 96.50 LR 97.18 97.20 TabSim 97.28 97.32 VI.

[REF2] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3
Title: UOBIT @ TAG-it: Exploring a Multi-faceted Representation for Profiling Age, Topic and Gender in Italian Texts
Chunk of text: Wa ∈ RM×SM, ba and bk are learnable parameters. The (*) T is the transpose operation and the output of the layer is O = [h0, ..., ht , ..., hN], a concatenation of the hidden states produced by the AttLSTM at each time step. As mentioned before, we propose a feature ensemble by using an interpretable multi-source fusion component (IMF). The IMF aims to combine features from different sources. A naive way of doing this is concatenating the vector representations into a single vector. This scheme considers all sources equally, but one source may yield a better result than others.

[REF3] - paperID: 4deed74a3eee7e629dce2b8ef1e437ca74b2e64a
Title: Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling Chunk of text: StdDev. .004 .005 .008 .003 .001 .001 determines the indexing throughput and query encoding latency, as well as the training batch size which influences the GPU memory requirements. The TREC-DL'20 query set was recently released, therefore most related work is missing results on these queries. We observe that the methods not using knowledge distillation and larger encoders (ANCE, LTRe) are outperformed on TREC-DL'19 by those that do use teachers (TCT, Margin-MSE), however on the sparse MSMARCO-DEV the result trend turns around. RocketQA

[REF4] - paperID: 44772b24ae2f68b77476c814b0607370f7195ddb
Title: Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval Chunk of text: , are of increasing interest, due to their use of the BERT embedding(s) for representing queries and documents. By using directly the BERT embeddings for retrieval, topic drifts for polysemous words can be avoided. To the best of our knowledge, our paper is the first work investigating PRF in a dense retrieval setting. Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval ICTIR '21, July 11, 2021, Virtual Event, Canada 3 MULTI REPRESENTATION DENSE RETRIEVAL The queries and documents are represented by tokens from a vocabulary V . Each token occurrence has a contextualised realvalued vector with dimension d, called an embedding. More formally, let f : V n → R $n \times d$ be a function mapping a sequence of terms {t1, . . .

[REF5] - paperID: 63483c9387d17e44eeb70c7321ad0dbb59b994fc
Title: Universal Multimodal Representation for Language Understanding Chunk of text: Recent studies have followed the two-path architecture [45, 46], in which the encoder consists of a joint embedding of textual and image representations extracted from both the images and corresponding caption. Notably, Engilberge et al. adopts RNN to encode sentence embeddings in the same space with extracted image representations from CNN. Portaz et al. enhances cross-modal retrieval using multilingual text. Inspired by the previous success of visual-semantic embeddings, we apply neural image retrieval from the joint space to fetch a group of associated images. 3 UNIVERSAL REPRESENTATION FRAMEWORK This section overviews our universal representation framework.

[REF6] - paperID: c537c75fa35d4060474862b82f31523811ae84da
Title: LawRec: Automatic Recommendation of Legal Provisions Based on Legal Text AnalysisChunk of text: (a) Transformer : it has achieved very good results in the field of machine translation. (b) SVM : it was first used to solve the two-classification problem in pattern recognition, and it has achieved good classification results in the fields of text classification, handwriting recognition, and image processing. (c) TextRnn : it is a model that uses RNN for text classification. (d) FastText : its biggest feature is that the model is simple, the training speed is very fast, and it is widely used in the field of text classification.

[REF7] - paperID: a609db40216a4071f9f739766c6691fa46fb8072
Title: Textual Representations for Crosslingual Information Retrieval Chunk of text: edu/˜kevinduh/a/wikiclir2018/ for the document indices. 4https://github.com/alvations/ sacremoses 5https://huggingface.co/Helsinki-NLP 6We use the opus-mt-en-de, opus-mt-en-fr, and opus-mt-en-jap models, their BLEU and ChrF scores (Papineni et al., 2002; Popovic´, 2015) can be found on https://huggingface.co/Helsinki-NLP (Tiedemann and Thottingal, 2020; Tiedemann, 2020) replicability of this paper. 3.1 Information Retrieval System We use the Okapi BM25 implementation in PyLucene as the retrieval framework with hyperparameter setting (k1 = 1.2, b = 0.75) (Manning et al., 2008). We consider the top 100 documents (topk = 100) in the search ranking as search results for each query. 3.1.1 Building index for the documents For each foreign language, we created an index for the documents with 5 TextField as follows: • id: the unique index of the document • surface: the raw text of the document • tokens: the document after tokenization •

subword: the document in SentencePiece subwords • char: the document in characters 3.1.2 Querying the document index During retrieval, each translated query is first processed into its respective text representations (tokens, subwords or characters) and parsed using Lucene's built-in query parser and analyzer. Additionally, we tried to improve the search results by combining and re-ranking the result sets from the different text representations. 3.1.3 Search result expansion Our intuition is that queries of more granular text representation can improve the robustness of the retrieval and potentially override the textual noise (e.g., misspellings are handled better for some languages).

[REF8] - paperID: bd23ce64a6422c1f73acf51675e53b7a06547da3
Title: UOBIT @ TAG-it: Exploring a Multi-faceted Representation for Profiling Age, Topic and Gender in Italian Texts
Chunk of text: Among the multi-lingual models, we investigate multilingual BERT (mBERT) (Pires et al., 2019) and XLM-RoBERTa (Conneau et al., 2020) which extends the corresponding pre-training over texts in more than 100 languages. Regardless of the adopted Transformer-based architecture, we also investigated the adoption of additional annotated material to support the training of complex networks over very short texts extracted from MEMEs. In particular, in Hate Speech Identification, we used an external dataset which addressed the same task, but within a different source. We thus adopted a dataset made available within the Hate Speech Detection (HaSpeeDe) task (Bosco et al., 2018) which involves the automatic recognition of hateful contents in Twitter (HaSpeeDe-TW) and Facebook posts (HaSpeeDe-FB). Each investigated architecture is trained for few epochs only over on the HaSpeeDe dataset before the real training is applied to the DANKMEMES material. In this way, the neural model, which is not specifically pre-trained to detect hate speech, is expected to improve its "expertise" in handling such a phenomenon (even though using material derived from a different source) before being specialized on the final DANKMEMES task3 . We trained UmBERTo both on HaSpeeDe-TW and on HaSpeeDe-FB and on the merging of these, too.

[REF9] - paperID: 65c2d2ffe45569101860a7defc7cccbd36b3602a
Title: Few-Shot Text Ranking with Meta Adapted Synthetic Weak SupervisionChunk of text: Moreover, they may often be overly confident and more unstable in the learning process (Qiao et al., 2019). A promising direction to alleviate the dependence of Neu-IR models on large-scale relevance supervision is to leverage weak supervision signals that are noisy but available at mass quantity (Zheng et al., 2019b; Dehghani et al., 2017; Yu et al., 2020). Through IR history, various weak supervision sources have been used to approximate querydocument relevance signals, e.g., pseudo relevance labels generated by unsupervised retrieval methods (Dehghani et al., 2017; Zheng et al., 2019b), and title-document pairs (MacAvaney et al., 2019). Recently, Zhang et al. (2020b) treat paired anchor texts and linked pages as weak relevance signals and propose a reinforcement-based data selection method ReInfoSelect, which learns to filter noisy anchor signals with trial-and-error policy gradients. Despite their convincing results, anchor signals are only available in web domains. Directly applying them to non-web domains may suffer from suboptimal outcomes due to domain gaps. To obtain weak supervision that adapts arbitrary domains, Ma et al.