ChatGPT

Topic Detection Techniques in Literature

TF-IDF with Linear Classifiers as Baselines

Term frequency–inverse document frequency (TF-IDF) features combined with linear models (e.g. logistic regression or linear SVM) have long served as strong baselines for topic detection and text classification [1] [2]. Despite their simplicity, such **bag-of-words models** can reach competitive accuracy when appropriately configured [3]. For instance, Wang and Manning (2012) showed that a *linear SVM* on TF-IDF features can outperform more complex methods on sentiment and topical datasets when using the right variants and features [4] [5]. Key enhancements include incorporating **n-grams** (especially bigrams) and applying suitable regularization. In fact, adding word bigrams to TF-IDF consistently improves performance on sentiment tasks and often boosts topic classification results (where unigrams alone capture most keywords) [6] [7]. Researchers note that using bigrams yielded better accuracy in many cases (whereas trigrams gave only minor or even slight negative returns) [8] [9].

Another successful trick is the *NB-SVM* feature reweighting proposed by Wang and Manning. In this approach, Naïve Bayes log-count ratios are used as feature values for a linear SVM, effectively blending generative and discriminative models [10] [4]. This simple **NB-SVM** variant was found to be “a strong and robust performer” across different text domains (sentiment, topic, subjectivity), often matching or surpassing more sophisticated methods [11] [12]. The NB-SVM’s interpolation parameter $β$ can be tuned to trade off NB vs. SVM confidence; values in the range 0.25–0.5 make the model **more robust**, with $β=0.25$ performing slightly better on short documents [13]. For binary multinomial NB (MNB) and NB-SVM, using **binary occurrence** features (presence/absence) instead of raw term counts yields a small accuracy gain (~1%) [14], echoing earlier findings that binarizing frequencies can improve stability [15].

In terms of hyperparameters, linear models with TF-IDF typically use **L2 regularization** (which proved more stable than L1 in text tasks [16]). Wang and Manning employed an L2-regularized SVM (LIBLINEAR implementation) and observed it worked best for these sparse features [16]. They set the SVM’s regularization parameter to around *C=0.1* in their experiments (with $C=1$ for the NB-SVM) [17], though in general C is tuned via cross-validation. Logistic regression performs very similarly to SVM in these high-dimensional TF-IDF settings [14], and indeed later studies continue to use **logistic or linear SVM as strong baselines**. For example, a 2016 Facebook research showed that “linear classifiers often obtain state-of-the-art performance if the right features are used” [18] [19] – confirming that a well-regularized TF-IDF + n-gram model is a tough baseline to beat. Recent benchmarks in specialized domains (e.g. legal text classification) still include TF-IDF+SVM as a first baseline, underscoring its robust performance across datasets [20] [2]. In summary, **TF-IDF with linear classifiers** – especially with enhancements like bigrams and NB-inspired reweighting – is well-supported in the literature as a performant and resilient topic classification technique.

Fine-Tuning XLM-RoBERTa for Topic Classification

Transformer-based language models have become dominant in NLP, and XLM-RoBERTa (XLM-R) is a multilingual RoBERTa variant that has been widely applied to topic classification across languages. Academic work demonstrates that fine-tuning XLM-R can achieve state-of-the-art results on topic detection tasks, often outperforming prior baselines. Conneau et al. (2020) report that XLM-R “significantly outperforms” the older multilingual BERT on cross-language benchmarks [21]. In particular, their 100-language XLM-R model boosted classification accuracy on low-resource languages by over 15% and even surpassed some monolingual models on certain tasks [21] [22]. For example, XLM-R obtained **5.1% higher** average accuracy on the XNLI language understanding benchmark compared to the previous state-of-the-art, while remaining competitive with strong monolingual models [23]. In large multilingual topic datasets, XLM-R consistently emerges as the top performer. A recent evaluation on 205 languages (SIB-200 corpus) found XLM-R to be the **best-performing pre-trained model** for topic classification in almost all languages tested [24]. Even for languages not seen in XLM-R’s pretraining, it generally outperformed other models as long as the script was familiar [24]. These results underscore that fine-tuned XLM-R is a *highly effective and robust choice* for topic detection, benefiting from its cross-lingual learned representations.

Fine-tuning XLM-R for classification requires selecting proper training strategies and regularization to generalize well. Researchers typically use the AdamW optimizer with a **learning rate on the order of 2×10^−5 to 3×10^−5** and include a small **weight decay (≈0.01)** to prevent overfitting [25] [26]. Empirically, these settings have been widely adopted as defaults in the literature, as they provide a stable balance between convergence speed and generalization [25]. For instance, one study tuned XLM-R on an emotion/topic dataset and chose a learning rate of 2e-5 with weight decay 0.01, 5 epochs, and batch size 8 – citing these as “standard fine-tuning practices” that yielded stable training and avoided overfitting on small datasets [27] [28]. A small batch (e.g. 8–16) can act as regularization when data is limited [27], and early stopping on validation loss is often used to curb over-training. XLM-R’s default dropout (around 0.1) and its **AdamW optimizer** already integrate regularization (via weight decay on transformer weights) which helps simplify the model and improve generalization [26] [29]. In sum, published fine-tuning recipes for XLM-R tend to agree on: a few epochs (3–5) with a **low learning rate (~2e-5)**, AdamW with **weight decay ~0.01**, and careful batch sizing and early stopping. When applied in this manner, XLM-R has delivered excellent topic classification performance, often establishing the new benchmark on both multilingual and monolingual topic datasets [24] [23]. Its strong results and capacity to handle diverse languages make it a go-to model in recent topic detection research.

LLMs (GPT) as Zero-/Few-Shot Topic Classifiers via Prompting

Recent research has explored using large language models (LLMs) like GPT-3 and its successors as *zero-shot or few-shot* topic classifiers through prompting. Instead of fine-tuning, these approaches supply the model with a prompt (e.g. a description of the task and/or a few example texts with their topics) and rely on the model’s in-context learning ability to predict the topic of new texts. Brown et al. (2020) famously demonstrated this capability with GPT-3, a 175-billion-parameter model [30]. In their work, GPT-3 was applied to a variety of classification tasks **without any gradient updates** – the tasks were specified purely via text prompts, and the model had to output the class. The results showed *surprisingly strong performance*: GPT-3 in a few-shot setting often **matched or approached state-of-the-art** accuracies from fully supervised models [30]. For instance, on news article topic classification (the AG News dataset), GPT-3 achieved about 90–91% accuracy with just a suitable prompt, whereas fine-tuned models reach ~95% [31] [32]. This gap narrows further with prompt engineering. Follow-up studies found that adding prompt techniques like *chain-of-thought reasoning* or retrieving similar examples can push LLMs even closer to supervised performance [33] [34]. In one benchmark, a prompt-tuned GPT-3 (with 16 example demonstrations) attained **96%+ accuracy on AG News**, essentially matching the fine-tuned RoBERTa baseline [35]. These findings validate that GPT-based models can serve as highly effective topic classifiers in a **zero or few-shot paradigm**.

Academic work has also elucidated *how* to prompt LLMs for classification. A common approach is to use a natural language template that turns the classification into a fill-in-the-blank or question-answering task. For example, Zhao et al. (2023) describe converting an input like *"The Warriors won the NBA championship 2022"* into a prompt such as: "The Warriors won the NBA championship 2022. **This topic is about [MASK]**." [36] [37] The language model (if it were BERT-like) or a generative LM (GPT-style) is then expected to predict the masked word or continuation – ideally outputting "sports" in this case, given its pre-trained knowledge [37]. Brown et al. used a similar idea but with GPT-3's autoregressive prompting: they prepend a few exemplars (e.g. "Article: … Topic: World News") and then prompt the model to complete "Topic:" for the new article [30]. Notably, this prompt-based classification requires no parameter update, leveraging the LLM's learned representation of topics. Researchers have shown that even **pure zero-shot** prompting (providing only an instruction or label names) can yield decent results, though formulating the prompt well is important [38] [39]. In cases where no training data is available, techniques like *verbalizer engineering* (choosing appropriate label words) or *automatic prompt generation* (to find good task descriptions) have been proposed to boost zero-shot accuracy [40] [41]. Overall, the literature (e.g. Brown et al., 2020 [30]; Kojima et al., 2022; Zhao et al., 2023) demonstrates that large GPT-style models can classify topics with impressive accuracy given the right prompts. This **prompt-based classification** paradigm has opened a new line of research into using LLMs as general text classifiers without task-specific training, achieving performance that in some cases rivals traditional supervised methods [32] [35]. These peer-reviewed studies and preprints underscore the promise of GPT variants as flexible zero-/few-shot topic detectors through carefully crafted prompting.

---

[1] [3] [18] [19] arxiv.org
https://arxiv.org/pdf/1607.01759

[2] Linear Classifier: An Often-Forgotten Baseline for Text Classification
https://www.researchgate.net/publication/372919038_Linear_Classifier_An_Often-Forgotten_Baseline_for_Text_Classification

[4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] Baselines and Bigrams: Simple, Good Sentiment and Topic Classification
https://aclanthology.org/P12-2018.pdf

[20] aclanthology.org
https://aclanthology.org/2022.acl-long.297.pdf

[21] [1911.02116] Unsupervised Cross-lingual Representation Learning at Scale
https://arxiv.org/abs/1911.02116

[22] [23] Unsupervised Cross-lingual Representation Learning at Scale
https://aclanthology.org/2020.acl-main.747.pdf

[24] SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects
https://arxiv.org/html/2309.07445v3

[25] Unsupervised Cross-lingual Representation Learning at Scale
https://www.researchgate.net/publication/343298778_Unsupervised_Cross-lingual_Representation_Learning_at_Scale

[26] [27] [28] [29] aclanthology.org
https://aclanthology.org/2025.semeval-1.247.pdf

30 [2005.14165] Language Models are Few-Shot Learners
https://arxiv.org/abs/2005.14165

31 32 33 34 35 tianweiz07.github.io
https://tianweiz07.github.io/Papers/23-emnlp.pdf

36 37 38 39 40 41 aclanthology.org
https://aclanthology.org/2023.acl-long.869.pdf