National University of Science and Technology
POLITEHNICA Bucharest
Faculty of Automatic Control and Computers
Computer Science and Engineering Department

DIPLOMA PROJECT

Analysis and classification of news in a semi-supervised manner

Ionescu Ștefan

Thesis advisors:

Prof. Dr. Ing. Mihai Dascălu
S.l. Dr. Ing. Ștefan Rușeți

BUCHAREST
2025

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In recent years, the online environment has grown exponentially, leading to an overwhelming volume of news and media content. Automatically classifying news articles is crucial for organizing and filtering this information. However, manually labeling large datasets is costly and time-consuming. This thesis examines a semi-supervised learning approach for Romanian news classification, utilizing both labeled and unlabeled data to enhance model performance.

We fine-tuned two transformer-based language models, such as RoBERTa-base and MT0, using techniques such as pseudo-labeling and data augmentation. The dataset consists of Romanian news articles categorized into thirty-two topics (politics, economy, sports, culture, events, meteo, and others). Our models achieved an accuracy of 0.82 and an F1 score of 0.81. These results demonstrate the effectiveness of semi-supervised learning for text classification in low-resource languages such as Romanian.

REZUMAT

În ultimii ani, mediul online a cunoscut o creștere exponențială, generând un volum copleșitor de conținut media și articole de știri. Clasificarea automată a articolelor de știri este esențială pentru organizarea și filtrarea eficientă a acestor informații. Totuși, etichetarea manuală a unor seturi mari de date este costisitoare și consumatoare de timp. Această lucrare explorează o abordare de învățare semi-supervizată pentru clasificarea știrilor în limba română, valorificând atât datele etichetate, cât și pe cele neetichetate pentru a îmbunătăți performanța modelelor.

Am antrenat două modele de procesare a limbajului natural de ultimă generație, RoBERTa-base și MT0, folosind tehnici precum pseudo-etichetarea și augmentarea datelor. Setul de date utilizat conține articole de știri în limba română, încadrate în treizeci și două de categorii tematice (politică, economie, sport, cultură, evenimente, meteo și altele). Modelele noastre au obținut o acuratețe de 0.82 și un scor F1 de 0.81 în clasificarea multi-clasă. Aceste rezultate demonstrează eficiența învățării semi-supervizate pentru sarcini de clasificare a textului în limbi cu resurse reduse, precum limba română.

1 INTRODUCTION

The rapid growth of digital media has transformed the way information is produced, distributed, and consumed. Every day, an immense volume of news content is published online across various platforms. This abundance of information makes it increasingly challenging to organize, categorize, and analyze news articles effectively. Automated news classification has thus become an essential tool for helping users and systems process and understand the vast amount of textual content available on the Internet.

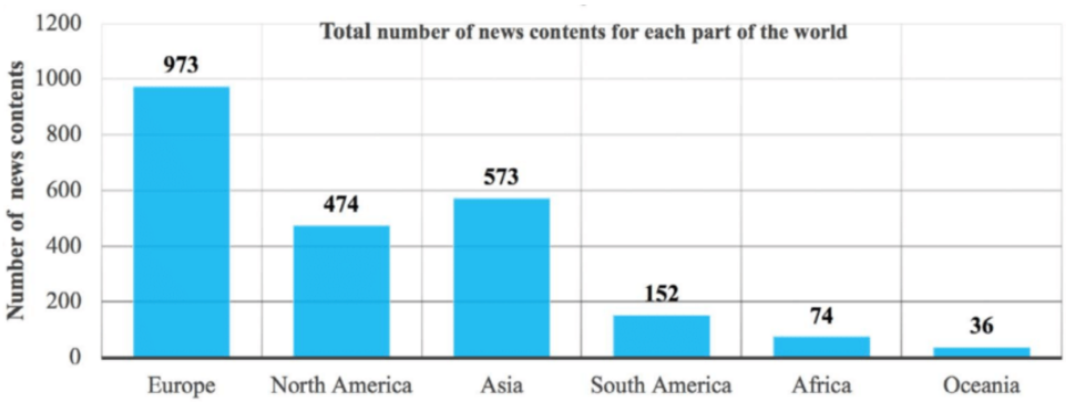


Figure 1: Ľuboš GREGUŠ et al. [1]

In this context, developing reliable methods for sorting and categorizing news articles into meaningful categories is a key goal in the field of Natural Language Processing (NLP). Recent advances in machine learning and deep learning have enabled the creation of more accurate and scalable models for this purpose.

1.1 Context

In an era where information spreads quickly on the Internet and social media, being able to sort and understand different types of news content has become very important and sometimes difficult. The large amount of online information can significantly influence how people think and see the world. In the past, classifying news was usually done with automated methods such as supervised learning.

However, one of the main problems with supervised learning is the necessity of a large amount of labeled data, which could be difficult to obtain and maintain. On the other hand, unsupervised learning does not offer us the desired performance for a viable detection.

A promising solution to this problem would consist of semi-supervised learning. The method combines the best of both worlds. It utilizes a limited number of labeled data points along with a large volume of unlabeled data to enhance the model's overall performance. Therefore, this thesis proposes to explore the efficiency and applicability of semi-supervised learning in detecting fake news, contributing to the development of more robust and scalable systems for combating misinformation.

1.2 Problem Statement

The main purpose of this thesis is to find a better way to automatically classify news articles without relying too much on large amounts of labeled data. Sorting news into the right categories is important for helping people understand what they're reading and for organizing information on websites, apps, and social media platforms. But doing this manually takes a lot of time, and labeling huge datasets for training models is often not realistic.

Traditional machine learning methods, like supervised learning, work well, but only when we have a lot of labeled examples. In many real-world cases, that type of data is hard to obtain. On the other hand, unsupervised methods don't need labels, but they often don't perform well enough to be useful.

This is where semi-supervised learning comes in. It combines a small set of labeled data with a larger set of unlabeled data to train models more efficiently. The goal of this project is to explore how this approach can help with news classification tasks and to see if it can offer a balance between performance and practicality. By doing so, this work hopes to support smarter and more scalable solutions for managing online content.

1.3 Objectives

The objectives of this thesis consist of the exploration of the potential of semi-supervised learning methods in improving the classification of news articles, particularly in situations where labeled data is limited or expensive to obtain. In real-world applications, gathering and manually labeling large datasets is time-consuming and often not feasible, especially when dealing with constantly evolving topics, multiple languages, or rapidly changing online content. Therefore, finding ways to reduce the dependence on fully labeled data is both a practical and necessary direction in machine learning research.

To address this challenge, the thesis investigates how different semi-supervised learning techniques can be applied to news classification tasks. The aim is to test whether models can achieve good performance by learning from a small labeled dataset, while also using a much larger pool of unlabeled news articles. The ability to make effective use of unlabeled data could significantly reduce the effort needed for annotation and improve model scalability.

The experimental part of this project is divided into two main phases. In the first phase, several semi-supervised approaches were tested using the BERT model, a widely used transformer-based model known for its strong performance on various natural language processing tasks. Techniques such as self-training, pseudo-labeling, and consistency regularization (or similar methods) were explored to identify which one provided the best results in the context of news article classification.

After evaluating the results from the BERT-based experiments, one of the tested methods was selected and applied to a second model, MT0. MT0 is a multilingual encoder-decoder model designed for general-purpose tasks, and testing it with the chosen semi-supervised strategy helped determine how well the approach generalizes across different types of models and architectures.

By comparing the performance of BERT and MT0 under the same semi-supervised learning method, this thesis aims to offer insights into model behavior in low-resource settings. The final goal is to identify practical techniques that can be used to build more efficient, scalable, and adaptable systems for automatic news classification. The findings of this work could be useful for organizations that need to manage large volumes of news content, such as media platforms, aggregators, or research institutions working on information filtering and topic detection.

1.4 Thesis Structure

The next chapter is divided into two main parts. The first part gives an overview of Natural Language Processing and language models. The second part focuses on semi-supervised news classification, explaining the task in general and reviewing existing datasets and related research. This chapter brings together information from previous studies.

The "Method" chapter describes how each language model was implemented, pointing out what makes each one unique. The fourth chapter shows the results of these implementations, using tables and images to make the data easier to understand.

The second-to-last chapter compares the two language models based on the results discussed earlier and also looks at the limitations of the project. Finally, the last chapter sums up the conclusions of the thesis and suggests possible improvements for future work.

2 STATE OF THE ART

Text classification represents the task of assigning predefined categories to textual inputs, which is a fundamental problem in Natural Language Processing (NLP). In the context of news articles, this task involves determining the topic, sentiment, or intent of a piece of text, such as classifying an article as "Politics," "Sports," or "Technology".

Modern methods for classifying text have significantly improved thanks to powerful language models like BERT, RoBERTa, DeBERTa, and T5. These models are first trained on huge collections of text from the internet, so they learn how the language works. After that, they can be adapted to specific tasks, such as news classification using only a small amount of labeled data. Because they understand the meaning and context of words in a sentence, they perform much better than older methods that only looked at individual words or word counts.

There are several common paradigms for applying Pretrained Language Models (PLMs) to text classification tasks:

- **Fine-tuning** - A pretrained model is adapted to the specific classification task by training it further on labeled data. For news classification, this typically involves adding a classification head and training on labeled articles.
- **Zero-shot classification** - Using generative models like T5 or MT0, one can prompt the model to classify text without any additional training.
- **Few-shot learning** - Instead of training on a large dataset, the model is prompted with a few examples per class (in-context learning). This is especially useful in scenarios with limited annotated data.
- **Chain-of-thought prompting** - Instead of directly asking for a label, the model is guided to reason step-by-step toward the correct classification. This method has shown improvements in tasks that require deeper understanding or multi-step reasoning.

Each of these strategies offers trade-offs between accuracy, computational cost, and data requirements. Fine-tuning generally offers the highest accuracy but requires labeled data and computational resources. Zero-shot and few-shot methods, on the other hand, are easier to use in new scenarios but may not perform as well in more complex classification tasks.

Importantly, recent advancements in generative Pretrained Language Models (PLMs), such as T5, FLAN-T5, and MT0, have helped close the distance between classification and generation. These models treat classification as a generation problem, where the output is simply the textual label. This unification simplifies architecture and enables multi-tasking across diverse NLP tasks.

In recent years, Semi-Supervised Learning (SSL) has emerged as a rapidly evolving field, largely driven by the increasing demand to develop high-performing machine learning models with minimal reliance on extensive labeled datasets. The high cost and effort associated with manual data annotation, particularly in domains requiring expert knowledge, have accentuated the need for methodologies that can effectively leverage large volumes of unlabeled data. SSL addresses this challenge by combining a small subset of labeled instances with a significantly larger pool of unlabeled data to improve model generalization.

A multitude of approaches have been proposed to exploit the latent information within unlabeled samples. Among the most prevalent are pseudo-labeling strategies, where labels are inferred for unlabeled data based on the model's current predictions. These pseudo-labels are then treated as ground truth in subsequent training iterations. While intuitive and straightforward to implement, naive pseudo-labeling is susceptible to propagating incorrect predictions, which may adversely affect the learning process. As a result, more recent studies have introduced mechanisms for filtering and weighting pseudo-labels according to measures of model confidence or prediction uncertainty, thereby enhancing the trustworthiness of the generated labels.

In parallel, consistency regularization has gained substantial attention as a complementary technique. This approach imposes a constraint that the model's predictions remain invariant under input perturbations, such as data augmentation, noise injection, or dropout. By encouraging the model to maintain consistent outputs for similar inputs, these methods facilitate the learning of smoother decision boundaries and contribute to improved robustness and generalization capabilities.

Moreover, the field has seen a notable shift toward adaptive and class-sensitive training paradigms. Traditional SSL frameworks typically employ uniform confidence thresholds and sample selection criteria across all classes, which can lead to suboptimal performance, particularly in imbalanced datasets. In contrast, recent advancements, such as Curriculum Pseudo Labeling (CPL) proposed in FlexMatch [7], dynamically adjust thresholds based on class-specific learning dynamics. This curriculum-inspired methodology allows the model to prioritize learning from simpler, well-represented classes before progressively incorporating more difficult or underrepresented ones, resulting in a more balanced and effective training process.

Another important line of research focuses on the quantification and integration of trust in pseudo-label selection. Techniques such as margin-based filtering, entropy-based scoring, and ensemble agreement have been employed to identify high-quality unlabeled examples, thereby mitigating the risks associated with low confidence or erroneous predictions. These trust-aware selection mechanisms serve to refine the pseudo-labeling process and reduce noise in the learning signal derived from unlabeled data.

Collectively, these developments reflect a broader trend toward more principled and data-efficient Semi-Supervised Learning (SSL) methodologies. By integrating pseudo-labeling, consistency regularization, uncertainty modeling, and adaptive selection strategies, contemporary

SSL frameworks have demonstrated the potential to rival fully supervised models, all while substantially reducing the burden of manual annotation. The following section surveys a selection of influential contributions from the literature that exemplify these innovations, with a focus on their mechanisms for data selection, trust calibration, and curriculum-driven learning.

In this work, we focus on classifying news articles, as it is a useful and important real-world task. This can mean telling the difference between informative news and opinions, or sorting articles by topic. Being able to classify news quickly and correctly is important for organizing information and suggesting relevant content to users. The next section shows how semi-supervised learning can improve these systems, especially when we don't have a lot of labeled data.

2.1 Language Models

2.1.1 BERT

Jacob Devlin et al. [2] propose BERT (Bidirectional Encoder Representations from Transformers). The model represents an important breakthrough in the way machines understand human language. Built by Google, BERT changes the way language models work by allowing them to look at words both to the left and to the right of a target word at the same time. This helps the model better understand the meaning of each word based on its full context in a sentence, something older models could not do as well because they only looked in one direction.

BERT is trained on huge amounts of text that are not labeled. During training, two tasks help the model learn. In the first task, called masked language modeling, BERT is shown sentences where some words are hidden, and it tries to guess the missing words based on the words around them. This teaches BERT to understand how words fit together in natural language. The second task is next sentence prediction, where BERT is asked to decide whether one sentence logically follows another. This helps the model learn how sentences connect and how meaning flows through a text.

What makes BERT especially useful is that after this pre-training, it can be fine-tuned to do many different language tasks, such as answering questions, understanding if two sentences mean the same thing, or finding names of people and places in text. Fine-tuning BERT is simple, meaning it only requires adding a small output layer and training on the new task. The core of the model, which has already learned a lot about language, remains the same.

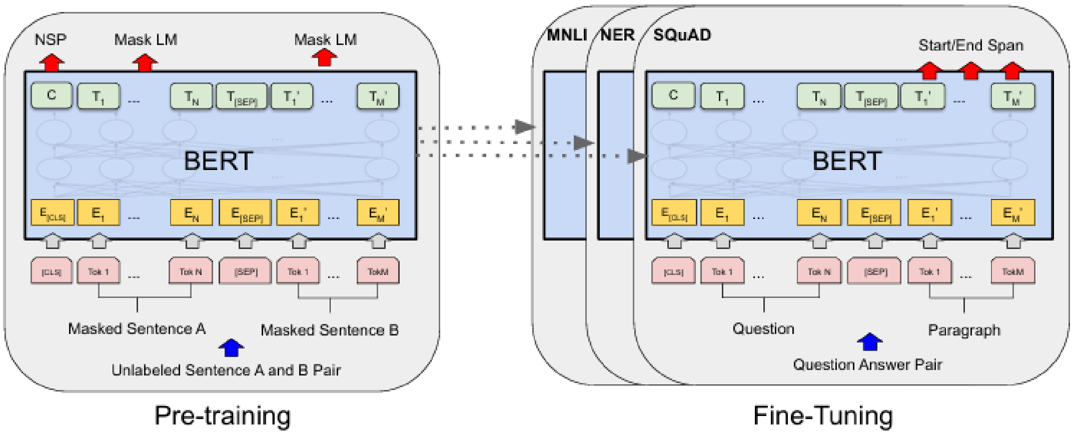


Figure 2: Jacob Devlin et al. [2]

The results BERT has achieved are very impressive. It has outperformed previous models on a wide range of language understanding tasks. On benchmarks like GLUE and SQuAD, BERT set new records for accuracy and quality. This is especially important because many tasks in natural language processing involve understanding complex relationships between words and sentences. The ability to use information from both directions gives BERT an edge in these tasks.

Another key advantage of BERT is that it works very well even when there is not much training data available for a new task. Because it has already learned so much from its large pre-training phase, it can generalize well to new problems with little extra training. This makes it a very practical tool for many real-world applications.

The introduction of BERT has had a big impact on the field of natural language processing. It has inspired many new models that build on its ideas and has pushed the boundaries of what AI systems can do with language. Today, BERT and its variations are used in many applications, from search engines to chatbots to document understanding systems. It represents an important step forward in helping machines better understand and work with human language.

2.1.2 MT0

Niklas Muennighoff et al. [3] propose the MT0 model. It is a multilingual extension of the T0 model, designed to improve how language models can generalize across different languages and tasks. It is based on the idea of multitask prompted finetuning, a process where the model is trained on many different tasks with natural language prompts to help it better understand how to follow instructions. While most earlier work in this area focused on English, this paper explores how the same approach can be applied to multilingual models.

The authors use two large multilingual models as a base: BLOOM and mT5. They first fine-tune these models on English-only tasks using the P3 dataset, which already helps them generalize to non-English languages thanks to the multilingual data present in the original pretraining. However, they go further and build a new dataset called xP3, which contains tasks in 46 languages, including translation, summarization, and even code generation. They also experiment with a version called xP3mt, where the prompts are machine-translated into each target language to better support non-English prompts.

Their experiments show that finetuning on English-only data already improves performance on non-English tasks, but using xP3 leads to even better results in both English and non-English. Interestingly, they also find that the models can generalize to languages they were not intentionally trained on, showing that the models learn higher-level capabilities that are not tied to any specific language.

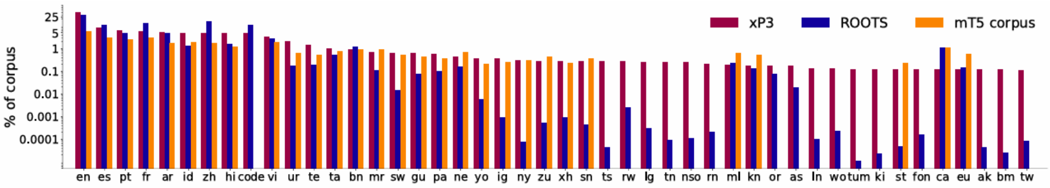


Figure 3: Niklas Muennighoff et al. [3]

When evaluating the models, the team notices strong improvements on many tasks, especially in natural language understanding tasks like coreference resolution, sentence completion, and natural language inference. They also observe that models finetuned with machine-translated prompts perform much better when evaluated on human-written non-English prompts. This highlights the importance of matching the language of the prompts to the target language of the task.

One interesting challenge they encounter is that finetuning on many short tasks biases the models toward producing short outputs, which can hurt performance on tasks that require longer responses, such as code generation or open-ended text generation. To mitigate this, they experiment with techniques like forcing a minimum output length during inference.

Overall, the MT0 paper demonstrates that multitask-prompted fine-tuning is a powerful method for enhancing the cross-lingual abilities of large language models. By carefully building multilingual datasets and prompts and applying this training method, the models can handle a wide range of languages and tasks with strong zero-shot performance, meaning they can perform new tasks without needing additional fine-tuning. The work also contributes useful resources to the community, making both the datasets and the trained models publicly available to support further research.

2.1.3 BERT vs MT0

BERT [2] and MT0 [3] are both powerful language models, but they were created with different goals in mind, and they use different training approaches. BERT was one of the first models to show that large language models can truly understand text by looking at it from both directions at the same time. It is trained by hiding words in sentences and asking the model to guess them, which helps it learn the meaning of words based on their full context. BERT is also trained to understand relationships between sentences. Once trained, BERT can be fine-tuned for many specific tasks, such as question answering or recognizing entities in text. However, BERT was originally built mainly for English and is not naturally suited for tasks in many different languages or for following complex instructions.

MT0, on the other hand, was designed from the start to be multilingual and instruction-based. Instead of just learning to predict missing words, MT0 learns by being asked to perform a wide variety of tasks using natural language instructions, such as "Translate this sentence" or "Summarize this article." The model is trained on many different tasks and in many different languages. A key idea behind MT0 is that by using this type of training, the model learns not just about language, but about how to follow instructions and generalize to new tasks it has never seen before. Another important point is that MT0 is trained on a special dataset called xP3, which includes 46 languages and many different kinds of tasks, including some that it wasn't even explicitly trained for.

While BERT works very well when it is fine-tuned for one specific task, it does not perform as well in zero-shot settings, meaning when it is asked to do something new without further training. MT0, thanks to its instruction-based training, is much better at this kind of generalization. It can often perform well even on tasks and in languages it was not explicitly trained on. Also, MT0 is better suited for handling full tasks like translation, summarization, or question answering in multiple languages, whereas BERT is more focused on understanding the structure of text.

In short, BERT was a breakthrough in teaching machines to understand the meaning of text, especially in English, while MT0 takes this further by teaching models to handle many languages and to follow human-like instructions across a broad range of tasks. Both models are important, but MT0 reflects how the field of language modeling has evolved to become more flexible and more useful in real-world settings where multilingual support and generalization are key.

2.2 Recent Advances in Semi-Supervised Learning

2.2.1 Multimodal Semi-Supervised Learning for Text Recognition

Aviad Aberdam et al. [4] propose a novel semi-supervised learning method tailored for scene text recognition, a task that involves reading text from images in natural environments like street signs or shop fronts. While synthetic data has long been the dominant resource for training models in this field, the authors argue that the abundance of real-world unlabeled images today offers a valuable alternative. However, making effective use of such data requires approaches that can learn reliably from both labeled and unlabeled examples.

The method, called SemiMTR, is built upon ABINet, a strong multimodal architecture that separates visual and language processing streams. The authors improve this architecture by introducing two key changes. First, they revise the pretraining of the visual component to include unlabeled data through a contrastive learning strategy. This helps the model better understand the structure of scene text by teaching it to distinguish between meaningful visual features (like characters) and irrelevant background noise. Notably, this is the first time contrastive learning has been successfully applied in this context, marking a significant step forward for scene text recognition.

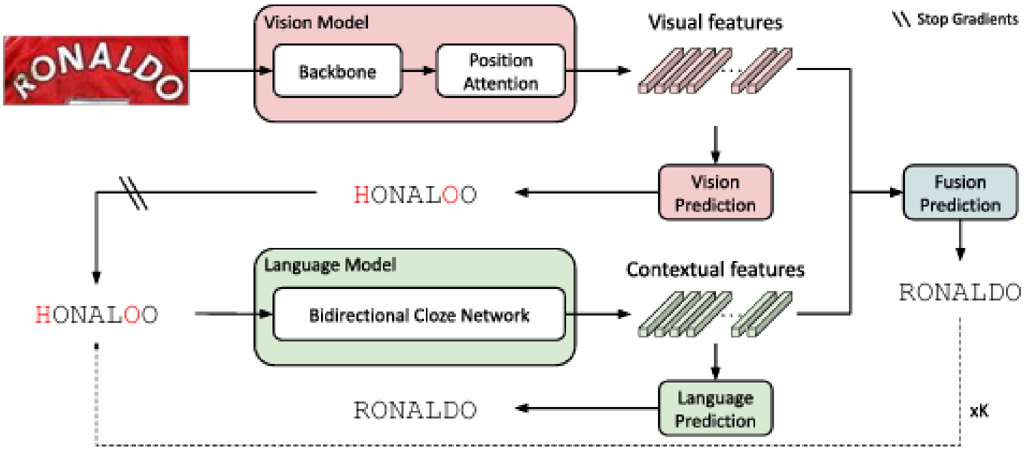


Figure 4: Aviad Aberdam et al. [4]

Second, they refine the final stage of training through what they call "multimodal self-teaching." Instead of relying on hand-crafted pseudo-labels or separate training cycles, the model teaches itself by enforcing consistency between predictions made on weakly and strongly augmented versions of the same input image. This character-level consistency is enforced across all three modules—vision, language, and fusion, making the model more robust and better aligned internally. It also avoids the need for extra training stages or auxiliary networks, keeping the system simple and efficient.

To evaluate the effectiveness of SemiMTR, the authors test it on several popular benchmarks, including both “common” and “non-common” datasets. The results show that SemiMTR not only outperforms previous semi-supervised methods but also surpasses fully supervised models that were trained on large synthetic corpora. Particularly impressive is its performance on non-common datasets, which include more diverse and realistic examples, proving its strength in generalizing to real-world settings.

Another important contribution is that the model achieves strong results even when trained only on real-world data, without any synthetic samples. When synthetic data is added, performance improves slightly, but the main gains come from the model’s ability to learn effectively from unlabeled real data. This represents a shift in thinking: instead of continuing to rely on large-scale synthetic datasets, the field can now start to focus more on leveraging the vast amount of naturally occurring images available online.

In essence, SemiMTR demonstrates that multimodal models, when paired with smart and lightweight semi-supervised strategies, can reach state-of-the-art results without the heavy cost of annotation or synthetic data generation. It presents a clear path forward for scene text recognition and opens new possibilities for SSL in multimodal learning beyond text recognition.

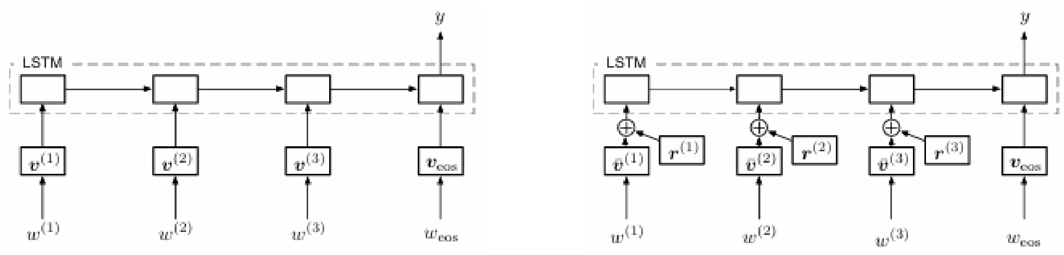
2.2.2 Adversarial Training Methods for Semi-Supervised Text Classification

Takeru Miyato et al. [5] made a significant contribution to semi-supervised learning (SSL) in Natural Language Processing (NLP) by introducing Virtual Adversarial Training (VAT), a regularization technique that improves model robustness by encouraging local smoothness in the decision boundary. The core idea behind VAT is to perturb inputs in the direction that would maximally alter the model’s predictions and then penalize that change, without relying on any labeled data.

Unlike traditional adversarial training, which typically aims to induce misclassifications, VAT is designed to enhance stability. It computes small perturbations in the input space, specifically in the direction that most disturbs the model’s output, and minimizes the KL divergence between the model’s original prediction and the one obtained after perturbation. Because this process does not depend on ground-truth labels, it is especially well-suited for semi-supervised tasks.

In Natural Language Processing (NLP), where inputs are discrete and non-differentiable, VAT applies perturbations in the embedding space rather than on the raw text. This allows the method to remain compatible with gradient-based training while benefiting from adversarial regularization. As a result, models trained with VAT become more robust to small changes and exhibit smoother, more generalizable behavior.

Experimental results showed that VAT achieved strong performance on benchmark datasets such as AG News and Yahoo! Answers, outperforming various baseline models in settings with



(a) LSTM-based text classification model.

(b) The model with perturbed embeddings.

Figure 5: Takeru Miyato et al. [5]

limited labeled data. Additionally, since VAT requires no changes to the model architecture and does not rely on language-specific tools, it can be easily integrated into a wide range of NLP pipelines.

Importantly, this work introduced perturbation-based regularization, well established in computer vision, into the realm of NLP. It bridged the gap between supervised robustness and unsupervised generalization, influencing later research in both domains.

In summary, VAT represents a principled and effective strategy for improving semi-supervised text classification. Its ability to regularize models without relying on labeled data makes it especially valuable in low-resource settings, and its impact continues to shape robust learning techniques today.

2.2.3 FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence

Kihyuk Sohn et al. [6] introduced FixMatch, a straightforward yet effective framework for semi-supervised learning. It combines consistency regularization with confidence-based pseudo-labeling, a method later adopted and extended in works such as FlexMatch (Bowen Zhang et al.[7]) and MarginMatch (Tiberiu Sosea et al.[8]).

The method works as follows. A weak augmentation is first applied to each unlabeled example, and the model's prediction is obtained. If the predicted class has a confidence score above a predefined threshold (commonly 0.95), that prediction becomes a pseudo-label. The same input is then passed through a strong augmentation, and the model is trained to match the pseudo-label. This process encourages the model to learn from unlabeled data by enforcing consistency across different transformations of the same example.

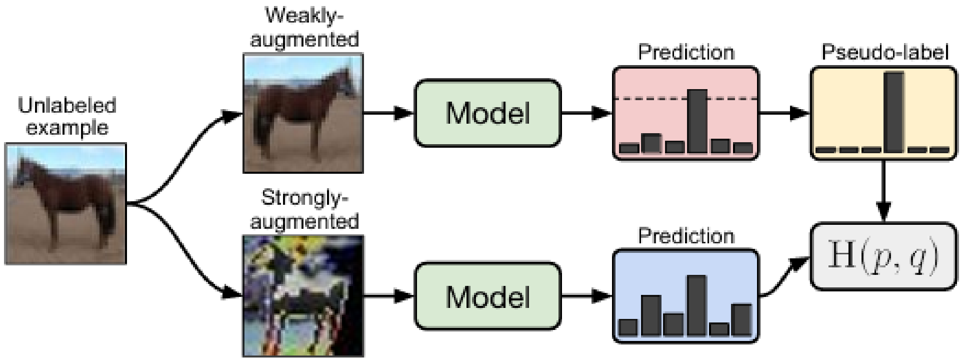


Figure 6: Kihyuk Sohn et al. [6]

One of FixMatch's greatest strengths lies in its simplicity. It does not require additional models, adversarial training, or uncertainty estimation. Despite its lightweight design, it achieves state-of-the-art results on several benchmark datasets, such as CIFAR-10, SVHN, and STL-10, even with very limited labeled data. The performance gains are largely due to the effective filtering of confident predictions and the use of strong augmentations that challenge the model to generalize.

FixMatch also had a broader impact on the direction of semi-supervised learning research. It demonstrated that well-designed heuristics, grounded in intuitive principles, can outperform more complex alternatives. Moreover, it laid the foundation for more adaptive techniques that followed, such as FlexMatch [7] and MarginMatch [8], which refine its core ideas to address issues like class imbalance and calibration.

In summary, FixMatch represents a clean and powerful approach to semi-supervised learning. Its ability to scale, its effectiveness in low-resource settings, and its influence on subsequent methods make it one of the most important contributions to the field in recent years.

2.2.4 FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling

Bowen Zhang et al. [7] introduce FlexMatch, which builds directly on the success of FixMatch [6] by introducing a more adaptive and class-aware approach to pseudo-labeling. While FixMatch applies a single confidence threshold uniformly across all classes, FlexMatch observes that not all classes are learned at the same pace. Some classes are easier for the model to recognize early in training, while others require more time and data.

To address this, FlexMatch introduces Curriculum Pseudo-Labeling (CPL), a dynamic strategy where confidence thresholds are adjusted individually for each class. Easier classes receive lower thresholds early on, allowing the model to generate pseudo-labels more readily. Harder classes begin with stricter thresholds, which are gradually relaxed as training progresses and the model becomes more confident. This curriculum-like structure mirrors how humans often learn, from simple concepts to more complex ones.

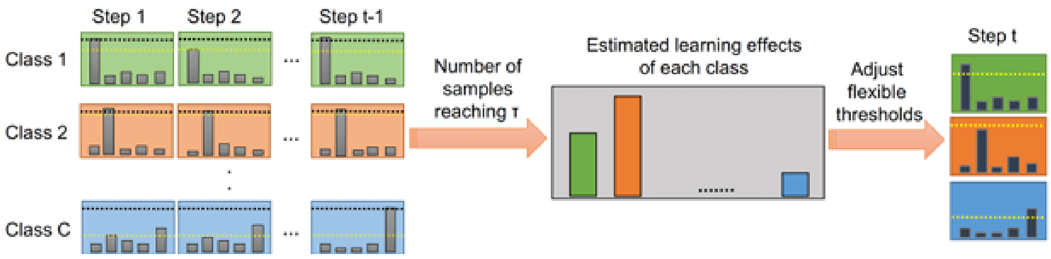

Step 1
Step 2
Step t-1
Class 1
Class 2
Class C
Number of samples reaching τ
Estimated learning effects of each class
Adjust flexible thresholds
Step t

Figure 7: Bowen Zhang et al. [7]

This adaptive design offers two main advantages. First, it helps filter out unreliable pseudo-labels for classes that the model has not yet learned well, improving stability during training. Second, it encourages more balanced learning across all classes, which is especially important in real-world datasets that often suffer from class imbalance.

FlexMatch achieves strong results on a wide range of image classification benchmarks, including CIFAR-10, CIFAR-100, SVHN, and STL-10. The improvements are particularly noticeable in challenging settings with limited labeled data or skewed class distributions. Despite these gains, FlexMatch maintains the efficiency and simplicity of FixMatch [6], requiring no significant changes to the model architecture or training process.

More broadly, FlexMatch represents a shift toward semi-supervised methods that are not only confidence-aware but also class-aware. It illustrates how curriculum-inspired techniques can guide the model to focus on reliable signals and progressively handle more difficult cases.

In summary, FlexMatch enhances the core ideas of FixMatch [6] with a thoughtful mechanism for class-specific learning. Adaptation to the model's learning progress and the structure of the data enables more effective and robust semi-supervised training.

2.2.5 MarginMatch: Improving Semi-Supervised Learning with Pseudo-Margins

Tiberiu Sosea et al. [8] introduce MarginMatch, a refined approach to pseudo-labeling by rethinking how model confidence should be measured when deciding whether to trust predictions on unlabeled data. While earlier methods, such as FixMatch [6] and FlexMatch [7], rely on confidence thresholds based on the highest softmax probability, MarginMatch proposes a more robust alternative: the pseudo-margin.

The pseudo-margin is defined as the difference between the top two class probabilities pre-

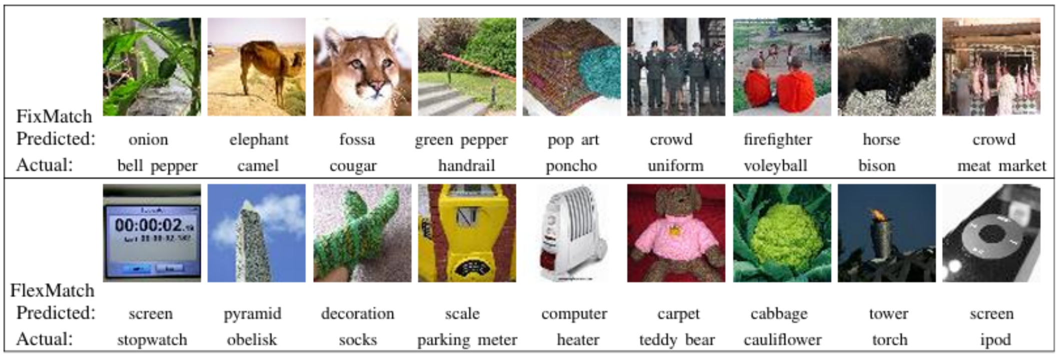


Figure 8: Tiberiu Sosea et al. [8]

dicted by the model. Instead of asking whether the model is confident in its top prediction, MarginMatch asks how much more confident it is in that prediction compared to the next most likely class. This subtle shift leads to a more stable and interpretable way to assess prediction certainty, especially in cases where softmax outputs are poorly calibrated.

During training, MarginMatch filters and weights unlabeled examples based on their pseudo-margins. Examples with large margins are more likely to be correct and are used more confidently in training, while those with smaller margins are either downweighted or discarded. This dynamic selection process helps the model avoid learning from ambiguous or noisy pseudo-labels, which is particularly valuable in low-resource scenarios.

MarginMatch integrates easily into the FixMatch [6] framework and retains its lightweight and efficient design. It does not require extra components or architectural changes, which makes it both practical and broadly applicable. Experimental results show that MarginMatch consistently outperforms previous methods, especially when labeled data is scarce or class distributions are imbalanced.

More broadly, MarginMatch reflects a growing trend in semi-supervised learning: moving beyond fixed thresholds and adopting more nuanced, data-driven measures of uncertainty. By leveraging the relative confidence between classes, it provides a more principled approach to filtering pseudo-labels.

In summary, MarginMatch is a simple yet effective enhancement to consistency-based SSL. Its strength lies not just in its robustness to noisy labels but in how it reframes confidence itself, not as a binary pass/fail but as a margin of certainty between plausible alternatives. By introducing pseudo-margins as a trust signal, it improves robustness in uncertain conditions and offers a reliable alternative to confidence-only strategies.

2.2.6 SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training

Hui Chen et al. [9] introduce SAT, or Simple Instance-Adaptive Self-Training, offering a lightweight yet powerful improvement to semi-supervised text classification. It revisits the idea of self-training and addresses a key limitation in earlier approaches, such as FixMatch [6], the reliance on manually defined weak and strong augmentations.

Rather than using fixed augmentation strategies, SAT adapts them on a per-instance basis. For each unlabeled input, the method generates two differently augmented versions. A small auxiliary model, called the augmentation choice network, is used to decide which of the two versions better preserves the semantic content of the original input. The selected version is treated as the weak augmentation and used to create a pseudo-label, while the other, more distorted version becomes the strong augmentation and is trained to match the pseudo-label. This enforces consistency while maintaining semantic relevance.

To train the augmentation choice network, SAT explores two strategies. One uses standard classification loss to estimate how close an augmented input is to the original. The other leverages cosine similarity and contrastive learning to rank augmentations based on how much meaning they preserve. Experimental results show that the latter provides slightly better performance, especially in low-resource settings.

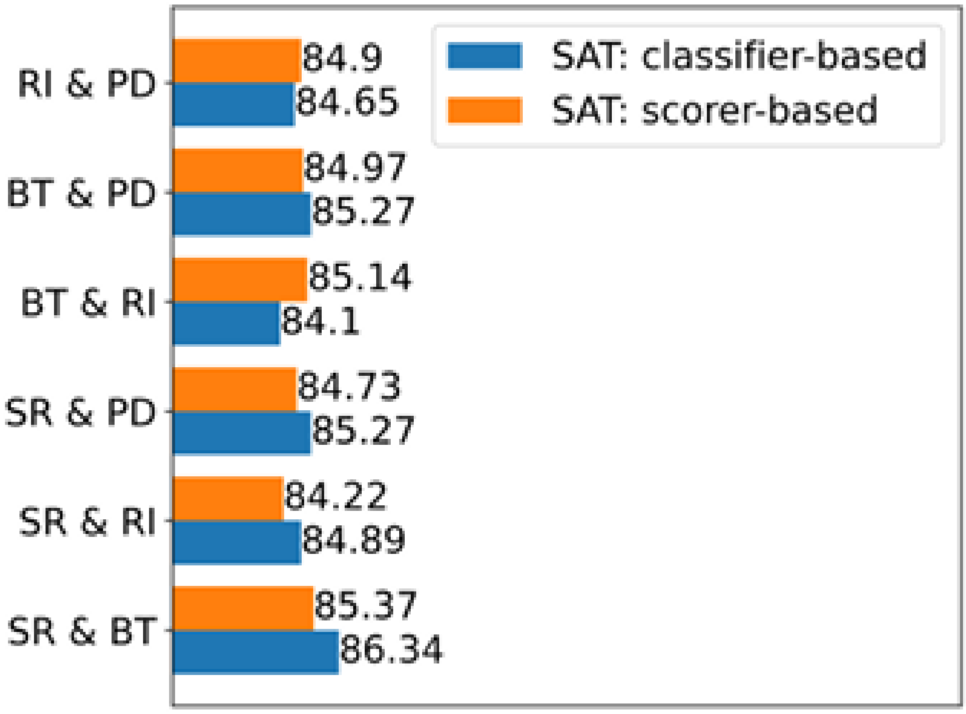


Figure 9: Hui Chen et al. [9]

SAT was evaluated on widely used datasets such as AG News, Yahoo! Answers, and IMDB. It consistently outperformed several strong baselines, including BERT, UDA, MixText, and FixMatch [6]. On average, it improved accuracy by 8.6% over BERT [2] and 2.6% over FixMatch, showing strong gains even when labeled data was scarce.

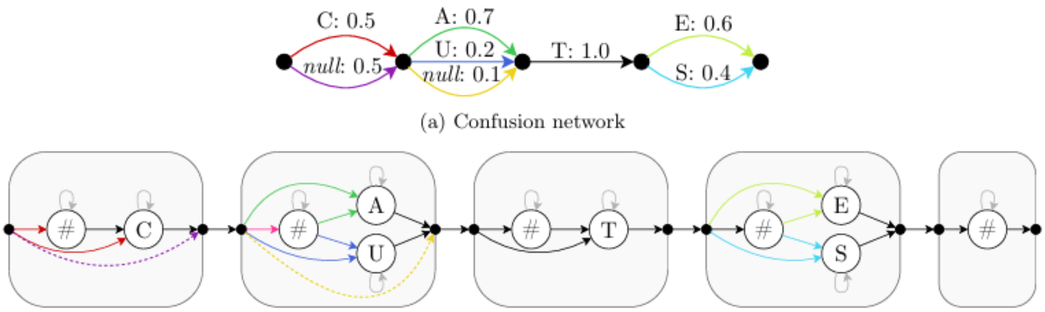
The method also includes ablation studies that highlight its robustness to different augmentation methods and its ability to scale with data. The best results were achieved using a combination of synonym replacement and back-translation.

In summary, SAT brings adaptability and semantic awareness to the augmentation process. By choosing augmentations based on how well they preserve meaning, it provides a more informed and reliable framework for self-training, especially in few-shot or low-data environments.

2.2.7 SoftCTC: Semi-Supervised Learning for Text Recognition Using Soft Pseudo-Labels

The SoftCTC method, proposed by Martin Kišš et al. [10], introduces a novel approach to semi-supervised learning for sequence tasks, particularly text recognition problems such as optical character recognition (OCR) and automatic speech recognition (ASR). Unlike traditional pseudo-labeling methods that rely on filtering based on fixed confidence scores, SoftCTC embraces prediction uncertainty by allowing the model to learn from multiple transcription hypotheses simultaneously.

At the core of the method is a modification of the standard Connectionist Temporal Classification (CTC) loss. Instead of using only the most confident sequence prediction, SoftCTC constructs confusion networks—structures that encode multiple plausible transcription candidates. The model then computes loss across this network, enabling it to consider several alternatives during training in a single forward-backward pass. This soft labeling mechanism reduces sensitivity to overconfident or incorrect predictions and leads to more stable learning.



(a) Confusion network

(b) Transcription confusion model. Black arrows, as well as the gray ones (self-loops), have the value of 1 while the colored ones have the value from the confusion network above. In confusion character groups, the dashed line represents the ϵ-transition.

Figure 10: Martin Kišš et al. [10]

One of the key advantages of SoftCTC is that it avoids the need for manually tuned confidence thresholds. It also eliminates the need to rerun CTC multiple times on different hypotheses, which improves training speed and scalability. This makes SoftCTC particularly appealing for settings with large amounts of unlabeled data and limited labeled samples.

SoftCTC was evaluated on challenging historical handwriting datasets, including Bentham, Rodrigo, and READ'16. In these experiments, the method achieved performance close to fully supervised models while using as few as 2,048 labeled lines. Benchmarks also showed that SoftCTC is computationally more efficient than multi-pass CTC alternatives, especially when deployed on GPUs.

In summary, SoftCTC provides an elegant and effective solution for semi-supervised learning in sequence recognition tasks. By using soft pseudo-labels and leveraging multiple plausible outputs, it offers both robustness and efficiency, making it well-suited for real-world applications where labeled data is scarce and sequence variability is high.

2.2.8 When and how epochwise double descent happens

Cory Stephenson et al. [11] explore a fascinating training dynamic known as epochwise double descent, where a model's generalization error first decreases, then increases, and finally decreases again as training continues over multiple epochs. This counterintuitive pattern has been observed in practice but is not yet fully understood, making it difficult to determine when to stop training to achieve optimal performance.

The authors propose a theoretical model that helps explain why and when this phenomenon appears. Their key insight is that certain features in the training data, while informative, are more difficult to learn. These "slow-to-learn" features interact with noise in the dataset, and this interaction can cause the generalization error to briefly worsen mid-training before recovering. This behavior disrupts the classical assumption that more training always leads to better generalization.

Interestingly, the paper shows that a specific level of label noise is necessary for double descent to occur. When noise is moderate, the model overfits to unhelpful signals before it fully captures the more meaningful but slower features, which creates a spike in the generalization error. However, if the noise becomes too high, the benefits of training begin to disappear entirely, and standard early stopping becomes more reliable again.

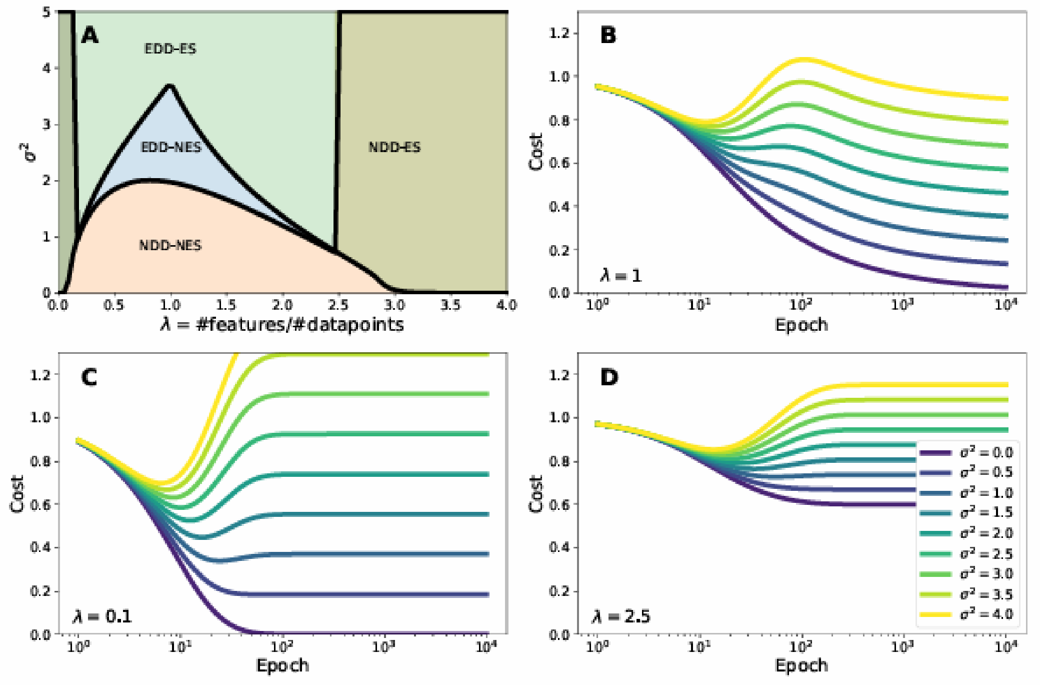


Figure 11: Cory Stephenson et al. [11]

To counteract double descent, the authors experiment with two strategies. One involves filtering out the problematic features that are learned too slowly, which stabilizes training but at the cost of losing valuable information. The other focuses on modifying the training dynamics, allowing the model to retain the useful features while avoiding the worst effects of overfitting. This second approach leads to better overall generalization than traditional training or aggressive early stopping.

Empirical results using deep neural networks support their theory, showing the same "rise-and-fall" behavior in test error across epochs. These findings not only help clarify a puzzling phenomenon but also provide practical guidance for managing training in settings where double descent is likely to appear.

In short, this work deepens our understanding of generalization dynamics in modern deep learning. It highlights how the timing and structure of learning, especially in the presence of noise, can significantly affect performance, and it points toward smarter training strategies that adapt to these behaviors.

2.2.9 Challenges in SSL

Even though semi-supervised learning (SSL) has come a long way in recent years, there are still several challenges that make it hard to use in real-world situations. Most of these problems are related to how we use unlabeled data and the assumptions that many SSL methods are based on.

Confirmation Bias

A common issue is that models can get stuck in their own mistakes. If a model makes a wrong guess on an unlabeled example and then treats that guess as the correct answer, it can reinforce the error and learn the wrong idea. Some methods, like FixMatch [6], try to avoid this by only keeping predictions that the model is very confident about, but being confident doesn't always mean being right. This issue is known as confirmation bias and has been studied in depth in works such as Arazo et al. [14].

Confidence Can Be Misleading

Many SSL methods trust predictions that have a high confidence score, usually the highest value from the model's output. But deep learning models often "feel" confident even when they're wrong [14]. This makes it hard to know when to trust a prediction. MarginMatch [8] improves this by not just looking at the top score, but also checking how much higher it is than the second-best option. Still, it's not a perfect solution.

Some Classes Are Harder to Learn

Not all categories are equally easy for the model to understand. Basic SSL methods use the same confidence rule for every class, which can lead to worse results on the harder ones. FlexMatch [7] fixes this by adjusting the confidence threshold for each class separately, starting with easier ones and slowly moving on to harder ones.

Text Augmentation is Complicated

In computer vision, Semi-Supervised Learning (SSL) works well because images can be easily modified (rotated, flipped, etc.). But for text, even small changes can alter the meaning of a sentence. Finding good and safe ways to augment text is still a big challenge, according to Markus Bayer et al. [16], although methods like SAT [9] attempt to select the best version of a modified sentence.

Unlabeled Data Can Be Messy

In practice, the unlabeled data you collect might come from a different topic or style than the labeled data. It can contain random or irrelevant examples. If the model learns from these, it could get worse. Some techniques, like SoftCTC [10], try to be more careful about uncertain data, but it's still hard to detect and avoid bad examples automatically [15].

Tuning Settings Can Be Tricky

SSL models often depend on settings like how confident a prediction needs to be or how strong the text augmentation should be. These settings can make a big difference, and they often change depending on the dataset. Finding the right values usually takes time and experience [7], [9].

Efficiency and Complexity

Some methods add extra tricks, like adversarial training or helper networks, to improve results [5]. These can help the model perform better, but they also make training slower and more complex. That can be a problem if we're working with limited computing power [15].

SSL is a powerful idea and can give great results, especially on standard datasets. But using it in real life is still tricky. Dealing with uncertainty, avoiding learning from bad examples, and keeping the training process simple and efficient are ongoing challenges that researchers continue to address.

2.2.10 Summary and Reflections

Semi-supervised learning (SSL) has progressed significantly in recent years, evolving from simple heuristics into a set of robust, adaptive strategies designed to make the most of the vast amounts of unlabeled data available today. What began as a straightforward idea—training models to learn from their confident predictions—has matured into a nuanced field that treats confidence, uncertainty, augmentation, and class imbalance as dynamic, learnable properties rather than static rules.

The foundation laid by models like BERT [2] demonstrated how large-scale pretraining on unlabeled corpora could capture deep contextual understanding of language. This paved the way for more flexible architectures like MT0 [3], which extended the paradigm by embracing multilingualism and instruction tuning, enabling strong cross-lingual generalization through prompt-based learning.

At the heart of many SSL methods lies pseudo-labeling. FixMatch [6] showed that even a simple approach—using high-confidence predictions from weakly augmented inputs as pseudo-labels for strongly augmented ones—can lead to remarkable performance, especially in image classification tasks. For instance, on CIFAR-10 with just 250 labeled examples, FixMatch achieved over 94 % accuracy, significantly outperforming previous SSL methods. Similarly, on datasets like SVHN and STL-10, it closed much of the gap between semi-supervised and fully supervised models. Although originally developed for computer vision, the underlying principle of consistency regularization combined with confident pseudo-labeling inspired many adaptations in natural language processing as well. However, applying a uniform confidence threshold across all classes exposed its limitations, especially when models learned some classes faster than others. FlexMatch [7] addressed this by dynamically adjusting thresholds for each class, following a curriculum that allows the model to focus on easier classes first and gradually take on harder ones. MarginMatch [8] took this idea further by measuring not just raw confidence, but the margin between the top two predicted class probabilities, resulting in more robust filtering of uncertain predictions.

Text augmentation—a central pillar of consistency-based SSL—presents unique challenges, as even small changes can drastically alter meaning. SAT [9] offered a novel solution by selecting augmentations adaptively, based on semantic preservation. This per-instance approach allowed the model to maintain consistency while avoiding harmful perturbations, especially valuable in few-shot scenarios where each labeled example matters.

Beyond classification, SoftCTC [10] pushed SSL into sequence modeling, particularly in text recognition. Rather than relying on a single hard pseudo-label, SoftCTC incorporated multiple plausible hypotheses using confusion networks, improving both robustness and learning stability. In a different vein, Virtual Adversarial Training (VAT) [5] introduced perturbation-based regularization to the NLP domain, encouraging models to make consistent predictions even when exposed to subtle adversarial noise in the input embeddings, without needing any labels at all.

Meanwhile, SemiMTR [4] extended SSL to the multimodal domain, leveraging consistency between visual and linguistic components in scene text recognition. It demonstrated that character-level self-training with real-world unlabeled images can outperform even fully supervised systems trained on synthetic datasets, pointing to the untapped potential of multimodal SSL.

On the theoretical side, the work on epochwise double descent [11] provided deeper insights into training dynamics, explaining how generalization error can rise and fall depending on when and how certain features are learned. This understanding helps practitioners better interpret model behavior during training and refine strategies like early stopping in SSL pipelines.

Taken together, these advances reflect a broader shift in the philosophy of SSL. Confidence is no longer binary; it is relational and context-sensitive. Pseudo-labels are no longer accepted blindly; they are filtered through adaptive, data-driven thresholds. Augmentation is no longer rule-based; it's learned. Generalization is no longer limited to known tasks or languages; it is enabled by instruction-tuning, multilinguality, and domain adaptation.

In essence, SSL is becoming more than a method—it is a learning mindset. One that embraces ambiguity, adapts to complexity, and leans into uncertainty rather than avoiding it. The techniques reviewed here collectively point to a future where models learn more like humans: gradually, flexibly, and always in context.

3 METHOD

In this chapter, we present the methodology used to address the task of Romanian-language news classification through both supervised and semi-supervised learning techniques. We begin by describing the datasets employed, which include a labeled corpus for initial training and evaluation, and a much larger unlabeled corpus used to enhance model performance through semi-supervised strategies. Following this, we introduce the two pretrained language models selected for our experiments—RoBERTa-based BERT and the multilingual encoder-decoder model MT0, highlighting their architectural differences and their suitability for this task.

We then detail the implementation of three semi-supervised learning methods: a standard pseudo-labeling pipeline, the more sophisticated FlexMatch [7] algorithm, and its improved variant, MarginMatch [8]. Each method is adapted to the specifics of text classification and carefully evaluated in terms of its contribution to model performance. Finally, we describe the evaluation metrics used to compare these approaches, focusing on both overall accuracy and class-level behavior using precision, recall, F1-score, and confusion matrices.

This methodological framework provides a systematic basis for analyzing how different architectures and training strategies influence the effectiveness of semi-supervised learning in a low-resource, multilingual setting.

3.1 Corpus

In this project, we used two main datasets: one labeled dataset for supervised learning and another unlabeled dataset for semi-supervised learning. Both datasets consist of Romanian-language news articles.

Labeled Dataset
The labeled dataset contains approximately 3000 samples, each labeled with one of 32 predefined categories, such as "Politică internă", "Meteo", "Sport", "Religie", "Cultură", "Lifestyle". Each sample includes the article title and its corresponding content. This dataset was used to train and evaluate the supervised baseline model as well as to provide a starting point for the semi-supervised approach.

Unlabeled Dataset
The unlabeled dataset consists of one million examples gathered from various online sources. These samples do not contain any category labels and were used as input for the semi-supervised learning pipeline. The goal was to leverage this larger pool of unlabeled data

to improve model performance through pseudo-labeling and consistency regularization techniques.

3.2 Proposed Solution

We used two state-of-the-art language models in our experiments: BERT and MT0.

BERT [2](Bidirectional Encoder Representations from Transformers) is a model based entirely on the Transformer architecture, more precisely on its encoder component. It differs from traditional models that process text in a single direction (left-to-right or right-to-left), as BERT uses bidirectional attention, reading the entire sentence simultaneously in both directions. This mechanism allows the model to better understand the meaning of each word in context. BERT comes in several sizes, such as BERT-Base (12 layers, 768 hidden units per layer, 12 attention heads) and BERT-Large (24 layers, 1024 hidden units, 16 attention heads), making it flexible for various tasks and resource constraints.

Alongside BERT, we also used MT0 [3], a multilingual encoder-decoder model based on the T5 (Text-To-Text Transfer Transformer) architecture. Unlike BERT, which is encoder-only and used mainly for understanding tasks, MT0 is a fully text-to-text model, capable of both understanding and generating text. It supports a wide range of languages, including Romanian, and is pretrained on large multilingual corpora with a unified objective: to generate target text from input text. MT0 also comes in different sizes, such as MT0-Small, MT0-Base, and MT0-Large, allowing trade-offs between model size and performance. Using both BERT and MT0 enabled us to compare the performance of encoder-only and encoder-decoder architectures in the context of Romanian news classification.

Since our datasets were in the Romanian language, we used a Romanian version of BERT, such as dumitrescustefan, to ensure the model was better suited to the linguistic characteristics of the language. This choice allowed for more accurate contextual understanding during both pretraining and fine-tuning stages, especially when handling Romanian-specific vocabulary and syntax. And for MT0, we used the base version of MT0 (MT0-Base) to balance model capacity and computational efficiency.

For the implementation, we used slightly different training strategies depending on the semi-supervised method. For FlexMatch and MarginMatch, we first trained the model in a fully supervised way using only the labeled data. This warm-up phase helped the model learn the structure of the Romanian text before moving on to the semi-supervised part.

For the Standard SSL approach, we didn't use a separate warm-up phase. Instead, we started directly with the semi-supervised loop, which included both supervised training on the labeled data and pseudo-labeling on the unlabeled examples from the beginning.

We applied all three methods: Standard SSL, FlexMatch, and MarginMatch, on the BERT model. Later, we also tested the Standard SSL approach on the MT0 model. We didn't apply FlexMatch or MarginMatch to MT0 because these methods are not well-suited for encoder-decoder architectures like MT0. Unlike BERT, which is a classification model, MT0 is a generative model that produces full text sequences as output. This makes it harder to use the same confidence-based or margin-based filtering strategies, which rely on classification probabilities. For that reason, we only used Standard SSL with MT0.

Each of these methods was implemented during the semi-supervised training stage, where the goal was to improve model performance by leveraging both labeled and unlabeled data in different ways.

Standard Semi-Supervised Learning:
In this approach, we organized the training loop into iterations. Each iteration began with supervised training on the labeled data, followed by a semi-supervised phase. In the semi-supervised part, the trained model was used to generate pseudo-labels for the unlabeled data, and only high-confidence predictions were retained and added to the training set. This process was repeated across multiple iterations to improve the model's generalization gradually.

FlexMatch:
The second method began with a fully supervised training phase over several epochs on the labeled data. After this pretraining stage, we applied the FlexMatch [7] algorithm in an iterative semi-supervised loop. Each iteration consisted of one training epoch and used both labeled and unlabeled data. We adapted FlexMatch, originally designed for image classification, to text classification by incorporating two types of textual augmentations: a weak transformation (a random word augmenter with a low transformation probability) and a strong one (the same augmenter with a higher probability). A central feature of FlexMatch is its dynamic confidence thresholding applied separately for each class: the model only uses pseudo-labels with confidence above the current threshold for that class, and the thresholds are updated after each iteration. The total loss combines supervised loss and unsupervised consistency loss.

MarginMatch:
The final approach we explored was MarginMatch [8], an extension of FlexMatch. While it builds on the same principles—consistency, regularization, and dynamic thresholds—its main innovation lies in how pseudo-labels are selected. Instead of using a fixed confidence threshold, MarginMatch considers the margin between the top two predicted class probabilities. At the beginning of the semi-supervised loop, a set of erroneous samples is created by randomly selecting unlabeled examples and assigning them to a virtual class. These samples help estimate a dynamic margin threshold, γ_t (gamma t), which acts as an additional filter. During training, a pseudo-label is accepted only if two conditions are met simultaneously: its margin exceeds both the class-specific dynamic threshold and the global γ_t. This dual filtering mechanism ensures that only the most reliable pseudo-labels contribute to training, effectively reducing noise as the model improves. Aside from this difference in label selection and filtering, the

rest of the training process remains similar to that of FlexMatch.

Through these three approaches, we aimed to evaluate the benefits and trade-offs of different semi-supervised learning strategies when applied to Romanian-language text classification.

3.3 Performance Metrics

To understand how well each method worked, we used a few common metrics that help us evaluate both the learning process and the final results. These included **accuracy, confusion matrix, precision, recall, and F1-score**. Each of them tells us something slightly different about how the model behaves.

Accuracy is the most straightforward — it simply shows how many predictions were correct out of the total. It's useful for getting a general sense of how well the model is doing overall.

Confusion Matrix provides a clear picture of how well the model is performing by showing the number of correct and incorrect predictions for each class. It helps identify specific patterns of misclassification, for example, which classes the model tends to confuse with one another. This visualization is beneficial for understanding the model's strengths and weaknesses beyond what accuracy or loss can reveal.

Precision tells us how many of the model's positive predictions were correct. It helps us understand how reliable the model is when it claims something belongs to a certain class. This metric is especially useful when making incorrect predictions can lead to confusion or unwanted consequences.

Recall shows how well the model manages to find all the relevant items. It measures how many of the actual, correct examples were successfully identified. This is important when missing something that should have been found is more serious than including a few incorrect results.

F1-score is a way to balance precision and recall. It combines both into a single value, which is helpful when you want to evaluate the model's overall ability to make correct and complete predictions, especially in situations where neither precision nor recall alone gives the full picture.

We compared the performance of different semi-supervised learning methods, like Standard SSL, FlexMatch, and MarginMatch, using the BERT model. We also applied Standard SSL to the MT0 model to see how it performs on a generative, text-to-text architecture.

By keeping the evaluation metrics the same, we were able to see how each method performs clearly. This helped us figure out not just which method worked best for BERT, but also whether those improvements carried over to MT0. In the end, MarginMatch gave the best results, especially when used with BERT, showing strong performance across all the metrics we tracked.

4 RESULTS

The figures provided illustrate how the F1-score evolves over training iterations for the BERT model under three different semi-supervised learning strategies: Standard SSL, FlexMatch, and MarginMatch. The first figure, corresponding to MarginMatch, shows an initial rise in F1-score followed by a slight drop, suggesting that the model quickly learns useful patterns before reaching a plateau. In contrast, the FlexMatch curve reveals a rapid and stable convergence; the model reaches a high F1-score early in training and maintains it across subsequent iterations. Meanwhile, the Standard SSL graph for BERT shows a steady decline in performance, which may indicate the model's struggle to benefit from pseudo-labels without any confidence-aware filtering. This drop suggests that pseudo-labeled examples may have introduced noise rather than reinforcing useful learning signals. In the case of MT0, the Standard SSL curve shows a quick rise followed by a plateau, suggesting that the model initially benefits from pseudo-labels but fails to improve further due to the lack of confidence filtering. MT0 appears slightly more robust to noise than BERT, but still stagnates in later iterations.

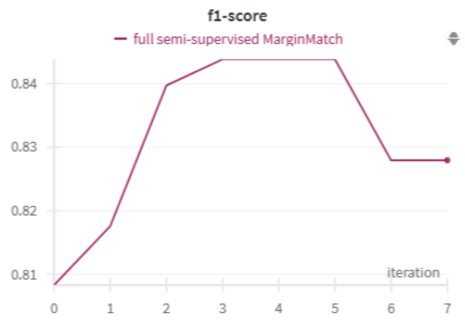


Figure 12: F1-Score Margin Match BERT

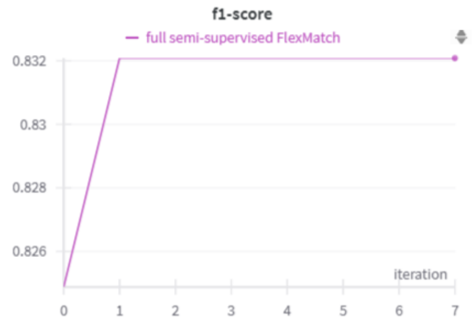


Figure 13: F1-Score Flex Match BERT

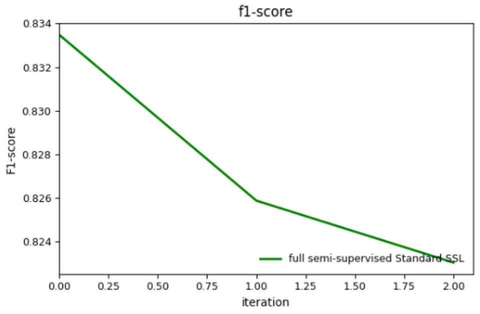


Figure 14: F1-Score Standard SSL BERT

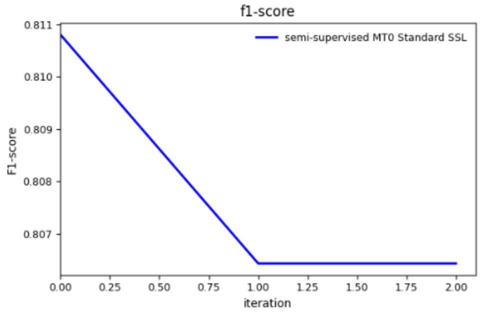


Figure 15: F1-Score Standard SSL MT0

These trends are confirmed by the performance metrics summarized in the first table. BERT trained with MarginMatch achieves the highest scores across all evaluation criteria—accuracy, precision, recall, and F1-score. With an F1-score of approximately 0.85, it outperforms the other approaches, demonstrating the effectiveness of filtering pseudo-labels based on the confidence margin between top predictions. FlexMatch also performs well, closely following MarginMatch, while Standard SSL lags, especially in F1-score, which drops to just over 0.83. This decline reinforces the idea that methods relying on fixed thresholds or naive pseudo-label selection can struggle in the absence of adaptive mechanisms.

Performance metric	**Margin Match**	**Flex Match**	**Standard SSL**
Accuracy	0.8495	0.83278	0.83612
Precision	0.84812	0.83255	0.83293
Recall	0.85069	0.83975	0.83023
F1-score	0.84382	0.83209	0.82034

Table 1: BERT

The second table presents the performance of the MT0 model when trained using only the Standard SSL approach. While the model performs reasonably well, reaching an F1-score of 0.806, it still falls short of BERT's results under the same conditions. This suggests that MT0, despite being a powerful multilingual encoder-decoder model, might not leverage pseudo-labeled data as effectively as BERT when lacking adaptive strategies such as FlexMatch or MarginMatch.

Performance metric	**Standard SSL**
Accuracy	0.80602
Precision	0.81377
Recall	0.8108
F1-score	0.80217

Table 2: MT0

Finally, the third table offers a direct comparison between BERT and MT0 using the Standard SSL method. BERT consistently outperforms MT0 across all metrics. The gap is particularly noticeable in accuracy and F1-score, where BERT maintains a slight edge. These results highlight how model architecture and pretraining objectives can impact the effectiveness of semi-supervised learning. In this case, BERT's encoder-only architecture, optimized for text understanding, appears to be better suited to classification tasks than MT0's more general-purpose sequence-to-sequence design.

Performance metric	**BERT**	**MT0**
Accuracy	0.83612	0.80602
Precision	0.83293	0.81377
Recall	0.83023	0.8108
F1-score	0.82034	0.80217S

Table 3: Standard SSL BERT vs MT0

5 DISCUSSION

5.1 Performance Comparison

To evaluate the effectiveness of each training method and model, we analyzed both quantitative performance metrics and the confusion matrix for the best-performing configuration. Table 1 summarizes the results obtained by BERT when trained with three different semi-supervised strategies: Standard SSL, FlexMatch, and MarginMatch. Among these, BERT with MarginMatch stands out. It achieved the highest values across all evaluation metrics: accuracy 0.8495, precision 0.8481, recall 0.8507, and F1-score 0.8438. These results indicate that the model was not only accurate in its predictions but also highly consistent across different classes. Precision and recall being closely balanced shows that the model managed to avoid both false positives and false negatives, which is crucial in multi-class classification tasks.

FlexMatch also performed strongly, with values just slightly below MarginMatch. It reached an accuracy of 0.8328 and a precision of 0.8326, while its recall was higher than Standard SSL. The F1-score, however, was lower than that of MarginMatch, suggesting that although FlexMatch maintained high confidence in its predictions, it may have been slightly less consistent when balancing false positives and false negatives.

Standard SSL, the most basic semi-supervised approach tested, achieved the lowest scores among the three, with an F1-score of 0.8203. While still functional, the method's lack of adaptive filtering or confidence-aware mechanisms made it less effective. Its lower recall, 0.8302, compared to MarginMatch implies that the model missed more relevant samples, possibly due to noise in pseudo-labeled data being incorporated too early or too indiscriminately.

These differences become even more evident when comparing performance across models. In Table 2, we present the results for the MT0 model trained only with the Standard SSL method. MT0 obtained lower scores across all metrics compared to BERT: an accuracy of 0.8110, a precision of 0.8132, a recall of 0.8169, and an F1-score of 0.8061. While these results confirm that MT0 remains a capable model, they also highlight the strength of BERT in this particular classification task, especially when paired with more refined SSL techniques like MarginMatch.

A direct comparison between BERT and MT0 using the Standard SSL method, as shown in Table 3, reinforces this conclusion. BERT surpasses MT0 in all four metrics, although with a moderate margin. This gap may be due to architectural differences: BERT's encoder-only structure is better suited for classification, while MT0's encoder-decoder design is generally optimized for text generation tasks. Additionally, BERT benefits more clearly from methods like MarginMatch, where class-wise filtering and margin-based selection help reduce noise in pseudo-labeling.

To complement the quantitative evaluation, we analyzed the confusion matrix for the configuration that performed best overall: BERT with MarginMatch. The figure below presents the confusion matrix, which visually illustrates how well the model distinguished between the 32 news categories. The results are compelling: correct predictions are tightly concentrated along the diagonal, indicating strong class-wise precision. This sharp diagonal shows that the model rarely confuses one category for another, even in cases involving semantically similar classes.

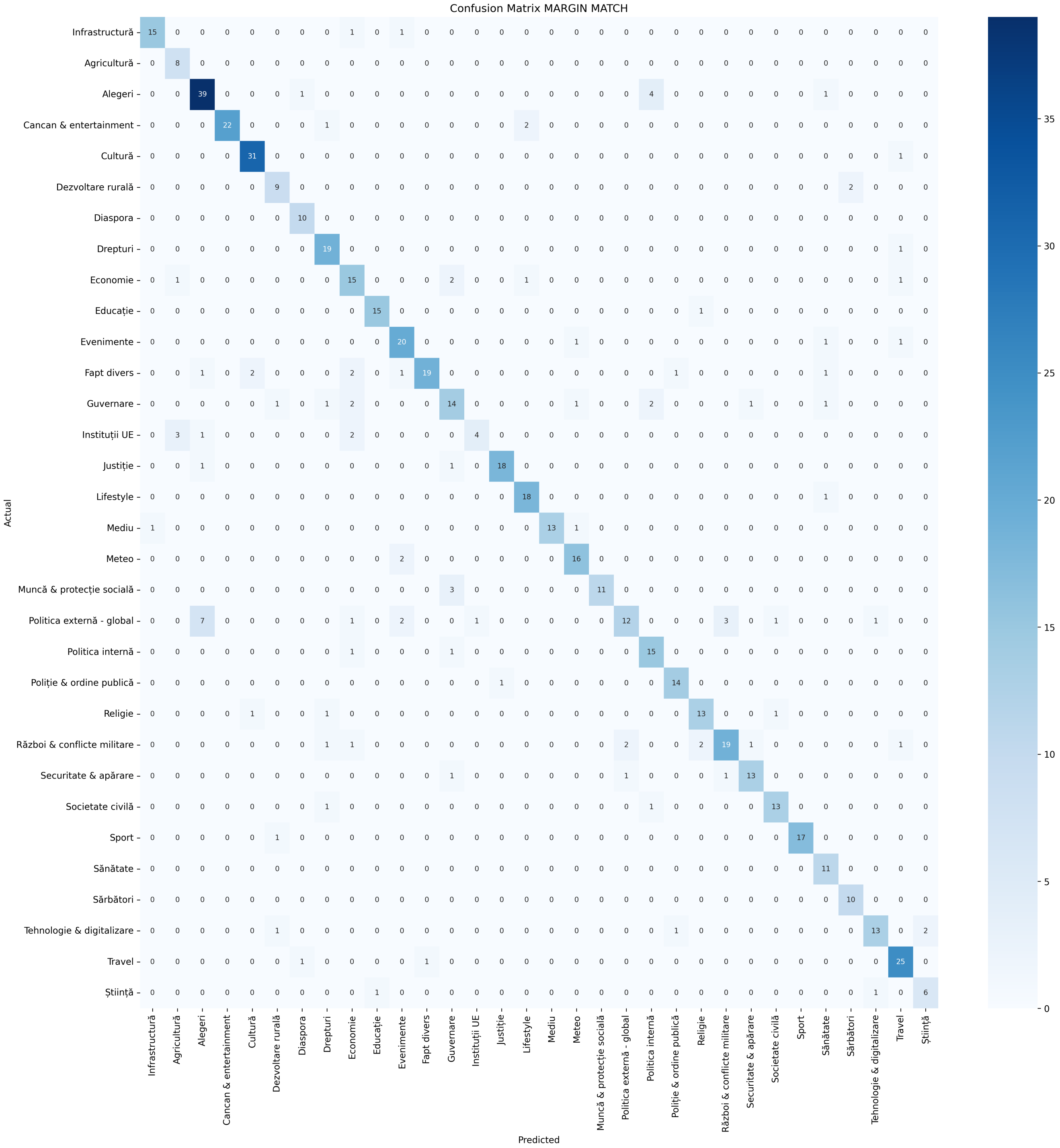


Figure 16: Margin Match Confusion Matrix BERT

Only a few off-diagonal values appear in the matrix, and even those are small. Most of these errors occur between conceptually related categories, such as “Lifestyle” and “Cancan & entertainment”, “Guvernare” and “Politica internă”, or ”Politică externă - global” and ”Alegeri”, where boundaries can be subtle. Although there are some overlaps between similar labels, the confusion matrix indicates that BERT with MarginMatch can still effectively distinguish between categories. This supports the high F1-score seen in the performance table.

Taken together, both the performance metrics and the confusion matrix suggest that BERT paired with MarginMatch is the most effective configuration for this semi-supervised multi-class classification task. It outperforms both Standard SSL and FlexMatch in terms of consistency, precision, and category-level clarity. FlexMatch remains a strong alternative, especially given its stability during training, but it lacks the margin-based refinement that gave MarginMatch the edge. Meanwhile, Standard SSL proves to be a solid baseline, but insufficient for achieving top performance without further improvements. MT0, while competent, appears less suited to this task when used with a basic SSL approach, underscoring the importance of aligning model architecture with the training strategy.

5.2 Limitations

Even though the results are promising, especially for BERT combined with MarginMatch, there are still some important limitations to keep in mind.

First, when looking at the confusion matrix, we can see that some categories still get confused with others quite often. This usually happens with topics that are semantically close or overlapping, like "Guvernare" and "Politica internă" or "Lifestyle" and "Cancan & entertainment". No matter how good the model is, when categories are similar in content or language, it's easy for the model to misclassify them. This suggests that the model may be picking up on surface-level patterns rather than deeper meanings.

Another limitation is the imbalance in the number of examples per category. Some classes, like "Cultură" or "Alegeri", have many samples, while others have very few. This naturally gives the model an advantage in the larger classes and can hurt its ability to learn the smaller ones well. We see this reflected in the confusion matrices, where smaller classes often show more prediction errors.

Lastly, although accuracy and F1-score are useful metrics, they don't always tell the full story, especially when dealing with many categories. Some classes might still be underperforming even if the overall scores look good, so it's important to look at per-class results and confusion matrices, not just the averages.

In summary, while the models perform well and show a lot of potential, they still face challenges with similar categories, class imbalance, and the limits of what overall metrics can reveal. These are important points to consider when thinking about future improvements.

6 CONCLUSIONS AND FUTURE WORK

This thesis explored the use of semi-supervised learning for classifying Romanian news articles, focusing on how to make better use of unlabeled data when labeled examples are limited. We tested three methods: Standard Semi-Supervised Learning, FlexMatch, and MarginMatch, using two language models: BERT and MT0. The results showed that semi-supervised approaches can significantly improve classification performance, even when the amount of labeled data is small.

Among the methods tested, MarginMatch performed the best, especially when combined with BERT. Its way of filtering examples using pseudo-margins helped reduce noise and made the model more stable during training. FlexMatch also showed good results, especially in handling class imbalance by adjusting confidence thresholds for each class. While MT0 showed strong performance with the standard method, it did not outperform BERT on this specific task.

Overall, our experiments confirm that semi-supervised learning is a practical and effective approach for text classification in low-resource languages, such as Romanian. It allows models to learn from large collections of unlabeled news, making them more useful in real-world scenarios where manual labeling is difficult or expensive.

There are several directions for future research. It would be useful to explore more advanced augmentation techniques tailored for the Romanian language. Augmentation plays a key role in improving generalization, and designing strategies that preserve meaning while introducing useful variation could boost performance further. We could look into combining semi-supervised learning with other strategies like active learning or human-in-the-loop systems. This would make it easier to build robust and adaptive models in environments where data is constantly changing.

REFERENCES

[1] Ľuboš GREGUŠ, Anna KAČINCOVÁ PREDMERSKÁ.
THE FOREIGN NEWS AND MEDIA IMAGE OF THE EUROPEAN UNION IN CURRENT TELEVISION NEWS PRODUCTION.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[3] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, MSaiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, AlhamFikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, Colin Raffel.
Crosslingual Generalization through Multitask Finetuning.

[4] Aviad Aberdam, Roy Ganz, Shai Mazor, Ron Litman.
Multimodal Semi-Supervised Learning for Text Recognition.

[5] Takeru Miyato, Andrew M. Dai, Ian Goodfellow.
Adversarial Training Methods for Semi-Supervised Text Classification.

[6] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, Colin Raffel.
FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence.

[7] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, Takahiro Shinozaki.
FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling.

[8] Tiberiu Sosea, Cornelia Caragea.
MarginMatch: Improving Semi-Supervised Learning with Pseudo-Margins.

[9] Hui Chen, Wei Han, Soujanya Poria.
SAT: Improving Semi-Supervised Text Classification with Simple Instance-Adaptive Self-Training.

[10] Martin Kišš, Michal Hradiš, Karel Beneš, Petr Buchal, Michal Kula.
SoftCTC – Semi-Supervised Learning for Text Recognition using Soft Pseudo-Labels.

[11] Cory Stephenson, Tyler Lee.
When and how epochwise double descent happens.

[12] Xiang Zhang, Junbo Zhao, Yann LeCun
Character-level Convolutional Networks for Text Classification.

[13] Yoon Kim.
Convolutional Neural Networks for Sentence Classification.

[14] Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, Kevin McGuinness.
Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning.

[15] Xiangli Yang, Zixing Song, Irwin King, Zenglin Xu.
A Survey on Deep Semi-supervised Learning.

[16] Markus Bayer, Marc-André Kaufhold, Christian Reuter.
A Survey on Data Augmentation for Text Classification.