Proposta recebida em Outubro 2024 e aceite para publicação em Dezembro 2024.

Estratégias de Seleção Informada de Dados para Aprendizado com Dados Escassos e Desbalanceados

Informed Data Selection Strategies for Few-Shot Learning on Imbalanced Data

Alexandre Alcoforado
Universidade de São Paulo

Lucas Hideki Takeuchi Okamura
Universidade de São Paulo

Thomas Palmeira Ferraz
Télécom Paris, Institut Polytechnique de Paris

Israel Campos Fama
Universidade de São Paulo

Bárbara Dias Bueno
Universidade de São Paulo

Bruno Miguel Veloso
Universidade do Porto, INESC-TEC

Anna Helena Reali Costa
Universidade de São Paulo

Resumo

A obtenção de dados anotados de alta qualidade é um dos principais desafios no Processamento de Linguagem Natural (PLN), especialmente em métodos de aprendizado supervisionado. Em cenários onde dados previamente anotados não estão disponíveis, soluções comuns como o *crowdsourcing* e a abordagem *zero-shot* frequentemente apresentam limitações, incluindo a necessidade de grandes volumes de dados e a falta de garantias quanto à qualidade das anotações. Tradicionalmente, os dados para anotação humana são selecionados de forma aleatória, uma prática que não só é custosa e ineficiente, mas também suscetível a viés, particularmente em conjuntos de dados desbalanceados, onde as classes minoritárias são sub-representadas. Para enfrentar esses desafios, este trabalho apresenta uma arquitetura de seleção automática e informada de dados, projetada para minimizar o volume de anotações necessárias enquanto maximiza a diversidade e representatividade dos dados selecionados. Entre os métodos avaliados, a Busca Semântica Reversa (RSS) se destacou, superando consistentemente a seleção por amostragem aleatória em cenários desbalanceados e melhorando o desempenho dos classificadores treinados. Além disso, realiza-se uma comparação entre a RSS e outros métodos baseados em agrupamento, discutindo seus pontos fortes e fracos.

Palavras chave

processamento de linguagem natural, seleção automática de dados, busca semântica reversa, dados desbalanceados, dados escassos, aprendizado com poucos dados

Abstract

Acquiring high-quality annotated data remains one of the most significant challenges in Natural Language Processing (NLP), especially for supervised learning approaches. In scenarios where pre-existing labeled data is unavailable, common solutions like crowdsourcing and zero-shot approaches often fall short, suffering from limitations such as the need for large datasets and a lack of guarantees regarding annotation quality. Traditionally, data for human annotation has been selected randomly, a practice that is not only costly and inefficient but also prone to bias, particularly in imbalanced datasets where minority classes are underrepresented. To address these challenges, this work introduces an automatic and informed data selection architecture designed to minimize the volume of required annotations while maximizing the diversity and representativeness of the selected data. Among the evaluated methods, Reverse Semantic Search (RSS) demonstrated superior performance, consistently outperforming random sampling in imbalanced scenarios and enhancing the effectiveness of trained classifiers. Furthermore, we compared RSS with other clustering-based approaches, providing insights into their respective strengths and weaknesses.

Keywords

natural language processing, automatic data selection, reverse semantic search, imbalanced data, low-resource, few-shot learning

1. Introdução

Em cenários da vida real, especialmente no campo de Aprendizado de Máquina (AM) em Processamento de Linguagem Natural (PLN), dados anotados são frequentemente recursos escassos e desafiadores de serem obtidos. Em muitos casos, pesquisadores e profissionais enfrentam a tarefa de desenvolver modelos precisos com um conjunto de dados anotados extremamente reduzidos ou até inexistentes. Para enfrentar esse desafio, o processo geralmente começa com a construção de um pequeno conjunto de dados anotados, utilizando-o como base para treinar modelos de AM através de métodos de aprendizado supervisionado. Em seguida, este processo pode ser iterado, criando conjuntos de dados anotados de tamanho crescente por meio de técnicas comumente referidas como Aprendizado Ativo (AA) (Ren et al., 2021).

Como uma abordagem alternativa para adquirir dados anotados, plataformas de *crowdsourcing* (ou

DOI: 10.21814/lm.17.1.451

anotação via participação coletiva) como *Amazon Mechanical Turk*[1] têm sido usadas nos últimos anos. No entanto, confiar exclusivamente em serviços de anotação humana dessas plataformas traz diversos desafios (Nowak & Rüger, 2010; Karpinska et al., 2021). A variabilidade na expertise entre os anotadores frequentemente resulta em critérios de anotação inconsistentes e, às vezes, em anotações conflitantes. Além disso, anotadores humanos podem encontrar dificuldades ao lidar com grandes conjuntos de dados, levando a erros e atrasos nos processos de anotação. Outra preocupação adicional reside na potencial introdução de viés pela subjetividade e preconceitos pessoais dos anotadores, o que pode afetar negativamente o desempenho dos modelos treinados. Para mitigar esses desafios, diversos trabalhos de pesquisa têm tentado resolver esses problemas, seja selecionando anotadores de alta qualidade em configurações de dados com múltiplas anotações ou empregando métodos diversos para ponderar a contribuição de cada anotador (Zhang et al., 2023a; Hsueh et al., 2009; Hovy et al., 2013; Basile et al., 2021).

Ainda assim, em ambientes com poucos recursos, é comum recorrer à amostragem aleatória de um subconjunto dos dados não anotados como estratégia para o processo de anotação (Tunstall et al., 2022; Beijbom, 2014). Essa abordagem envolve a seleção aleatória de alguns exemplos, que são então anotados para formar o conjunto de dados de treinamento inicial. No entanto, essa metodologia também pode ser subótima, pois negligencia as características específicas dos dados e os requisitos do modelo que será treinado. Em outras palavras, dados amostrados aleatoriamente podem não representar adequadamente todo o espectro de classes ou conceitos presentes no conjunto de dados e, por consequência, prejudicar o desempenho do modelo.

O advento de métodos de classificação *zero-shot* (ou classificação sem amostras de referência) trouxe uma abordagem prática para realizar anotações iniciais sem depender de dados de treinamento previamente anotados. Embora amplamente estudados por diversas comunidades de pesquisa, como exemplificado por Alcoforado et al. (2022) no contexto da língua portuguesa, esses métodos historicamente apresentaram desempenho inferior em relação aos métodos de classificação *few-shot* (ou classificação com poucos exemplos de referência). Recentes avanços no Processamento de Linguagem Natural, impulsionados pelo surgimento de Grandes Modelos de Linguagem (LLMs) de propósito geral, como os introduzidos por Brown et al. (2020) e Touvron et al. (2023), ampliaram as possibilidades de aprendizado multitarefa e resolução de problemas *zero-shot* (Ferraz et al., 2024a). Apesar de seu notável potencial, esses modelos ainda enfrentam dificuldades em se adaptar a domínios específicos, onde conhecimentos altamente especializados podem estar completamente ausentes de seus dados de treinamento, como apontado por Yang et al. (2023), Ferraz et al. (2024b) e Zhang et al. (2023b).

No âmbito da classificação de texto *few-shot*, o desafio de adquirir dados anotados torna-se cada vez mais difícil, especialmente quando confrontado com conjuntos de dados desbalanceados (Ferraz et al., 2021). Conjuntos de dados de *benchmark* comuns usados para tarefas de classificação de texto *few-shot* frequentemente exibem um equilíbrio ou apenas um ligeiro desbalanceamento na proporção do número de exemplos (instâncias) pertencentes a cada classe no conjunto de dados. No entanto, esses conjuntos de dados representam exceções raras em um cenário prático, onde distribuições de dados são tipicamente enviesadas e desbalanceadas, refletindo a complexidade inerente do mundo real. A prevalência de dados desbalanceados representa um desafio significativo, pois estratégias tradicionais de amostragem aleatória tornam-se cada vez mais subótimas. Em cenários onde uma classe domina esmagadoramente, a amostragem aleatória tende a favorecer a classe majoritária, resultando em uma seleção de dados que representa inadequadamente as classes sub-representadas e raras. Um caminho viável é a utilização de algoritmos mais sofisticados como amostragem estratificada, que assume a existência de estratos na distribuição. Entretanto, não há garantias de que esses algoritmos sejam mais eficientes em cenários de ausência de dados anotados, uma vez que os estratos precisam ser encontrados de maneira não supervisionada.

As próprias tarefas de classificação de texto *few-shot* e *zero-shot* são subconjuntos específicos de uma tarefa chamada *X-shot learning* (Xu et al., 2024), onde $X \in [0, +\infty[$ para cada classe. Essa definição de tarefa abrange uma quantidade muito maior de aplicações práticas, onde humanos costumam ter muitos exemplos de algumas classes, poucos de outras e nenhum de outras. No domínio dos métodos de aprendizado *X-shot*, a necessidade de atingir o balanceamento é eliminada, o que geralmente tem um custo no desempenho do modelo em alguns aspectos. A possibilidade de utilizar números diferentes de exemplos anotados por classe é promissora, mas foge do escopo desse trabalho, uma vez que modelos preparados para essa tarefa ainda são escassos ou têm desempenho inferior a métodos *few-shot*, ainda que com poucos exemplos.

[1] https://www.mturk.com/

Para enfrentar os desafios citados, neste artigo introduzimos uma arquitetura inovadora de seleção automática de dados para aprendizado *few-shot*. Nossa abordagem é projetada para identificar os dados mais informativos e representativos que devem ser anotados por humanos em cenários com poucos recursos e escassez de anotações. Ela aproveita uma estrutura que ordena sistematicamente os dados com base na sua probabilidade de (i) pertencerem a classes distintas, evitando assim redundâncias desnecessárias nos esforços de anotação humana, e (ii) melhorarem o desempenho geral do modelo de aprendizado. Nossa avaliação dessa abordagem abrange diversos conjuntos de dados de processamento de linguagem natural com poucos recursos, demonstrando sua capacidade de minimizar redundâncias nos esforços de anotação humana e melhorar o desempenho do modelo em comparação com estratégias tradicionais de amostragem aleatória ou seleção manual de dados, especialmente em casos com um número limitado de exemplos anotados.

Em resumo, este trabalho apresenta duas contribuições principais:

1. A introdução de uma arquitetura de seleção automática de dados para aprendizado *few-shot* que aproveita os princípios do aprendizado ativo para identificar os dados mais informativos e representativos para anotação.
2. Uma análise extensiva de várias implementações de nossa arquitetura, destacando sua eficácia e eficiência na construção da primeira versão de um conjunto de dados no contexto da classificação de texto com poucos recursos.

Nossos resultados enfatizam os benefícios da seleção informada de dados, que não apenas simplifica o processo de anotação, mas também resulta em um conjunto de dados anotados mais diversificado. Além disso, modelos treinados com esses conjuntos de dados diversos exibem desempenho aprimorado, o que pode beneficiar iterações subsequentes do conjunto de dados com técnicas de Aprendizado Ativo. Nossos experimentos revelam o potencial das estratégias de seleção informada de dados em enfrentar os desafios do aprendizado *few-shot* em cenários de PLN com poucos recursos.

As partes seguintes deste documento estão organizadas da seguinte maneira: a seção 2 discute os trabalhos relacionados à aprendizagem com dados desbalanceados, dividindo-os em duas categorias: aqueles que não abordam diretamente o desbalanceamento e aqueles que tratam explicitamente dessa questão. Nosso método se insere na segunda categoria, ao oferecer uma abordagem informada para seleção de dados, projetada para lidar de forma eficiente com conjuntos desbalanceados. A seção 3 introduz os métodos propostos e formaliza matematicamente os procedimentos de seleção. A seção 4 apresenta os detalhes dos experimentos realizados, incluindo os dados utilizados e a implementação dos métodos. Na seção 5, discutimos os resultados obtidos, e na seção 6, destacamos as conclusões do trabalho e sugerimos direções para trabalhos futuros.

2. Aprendendo com Dados Desbalanceados

Nesta seção, apresentamos e discutimos trabalhos relacionados, organizados em duas categorias principais. A primeira abrange estudos que não requerem o balanceamento dos dados para o aprendizado supervisionado, focando em cenários em que o desbalanceamento nos dados de treinamento anotados é compatível com o algoritmo utilizado no treinamento. Na segunda categoria, exploramos métodos que impõem a necessidade de balanceamento no aprendizado supervisionado, contexto no qual nossas propostas se configuram como soluções viáveis.

2.1. Desbalanceamento no Treinamento

A ausência de balanceamento em conjuntos de dados de treinamento é um desafio amplamente estudado no campo de PLN, com diversas abordagens propostas para lidar com essa característica sem recorrer a um balanceamento artificial dos dados. Essas estratégias reconhecem que é possível explorar de forma eficiente a informação presente nos dados desbalanceados, desde que haja uma escolha cuidadosa dos métodos de seleção e treinamento.

O aprendizado ativo (AA) é uma família de estratégias, priorizando a seleção de amostras que trarão o maior benefício para o modelo com o mínimo de anotações (Ren et al., 2021). Este enfoque é valioso quando as anotações são caras ou difíceis de obter — como é frequentemente o caso. A amostragem por incerteza é uma técnica comum dentro do aprendizado ativo, em que um modelo é usado para identificar instâncias de baixa confiança (Zhu et al., 2010). No entanto, um ponto negativo dessa abordagem é que ela pode ser míope, focando em exemplos marginais que podem não ser representativos do conjunto de dados como um todo.

Outra técnica de aprendizado ativo é a consulta por comitê, que utiliza múltiplos modelos para identificar exemplos com uma grande discordância entre previsões (Kee et al., 2018). Apesar de ser mais robusta à variabilidade e ao ruído nos dados, ela pode ser computacionalmente cara e complexa para gerenciar, especialmente com um grande número de modelos ou classes.

As amostragens por margem e por entropia consistem em técnicas que tentam identificar exemplos onde o modelo está incerto entre várias classes (Ducoffe & Precioso, 2018; Li et al., 2011). Essas abordagens podem ser eficazes em identificar exemplos limítrofes, mas podem falhar em capturar a diversidade dos dados, concentrando-se excessivamente em áreas de alta incerteza.

Métodos de aprendizado ativo que endereçam a diversidade dos dados, como os propostos por Sener & Savarese (2018) e Zhang et al. (2021), buscam evitar o viés de seleção ao garantir que o conjunto anotado inclua uma ampla gama de variações nos dados. No entanto, uma desvantagem dessas abordagens é que elas podem ignorar a utilidade prática de cada exemplo, resultando na seleção de dados que são diversos, mas sem garantir que sejam os mais informativos para o modelo.

O *X-shot learning*, onde X representa a frequência dos rótulos, oferece uma alternativa mais flexível, permitindo que o número de exemplos anotados por classe varie (Xu et al., 2024). Este tipo de abordagem tenta adaptar o modelo para funcionar com diferentes quantidades de dados disponíveis por classe. No entanto, pode ser desafiador para os modelos lidarem com a grande variação no número de exemplos, e pode haver uma perda de desempenho em classes com pouquíssimos ou nenhum exemplo.

Apesar de o BinBin (Xu et al., 2024) demonstrar avanços significativos em *X-shot learning*, o modelo ainda sofre com a necessidade de lidar com um grande número de classes e a complexidade associada à criação de instruções para guiar a classificação. Além disso, a aplicabilidade em domínios específicos onde o vocabulário e os conceitos são altamente especializados permanece uma limitação a ser superada.

Embora a otimização direta do esforço humano na anotação de dados seja um objetivo altamente desejável, observa-se, com base nos trabalhos citados, que quantificar e integrar esse fator de forma sistemática na seleção de amostras apresenta desafios significativos. Estratégias que combinam incerteza e diversidade visam promover uma seleção de dados mais equilibrada; no entanto, sua implementação frequentemente envolve complexidades, e nem sempre garantem boa representatividade ou informatividade dos exemplos selecionados. Além disso, é fundamental reconhecer que, embora certas abordagens possam ser eficazes em contextos específicos, elas não são universalmente aplicáveis devido às suas limitações inerentes. A escolha da estratégia mais adequada deve considerar não apenas a natureza do desbalanceamento dos dados, mas também características intrínsecas desses dados, os objetivos específicos do modelo, o custo associado às anotações e as restrições de recursos computacionais.

2.2. Balanceamento no Treinamento

Diferentemente das abordagens mencionadas anteriormente, existem métodos que consideram o balanceamento dos dados como um pré-requisito essencial para o treinamento eficaz de modelos de aprendizado de máquina. Essas técnicas, amplamente utilizadas no aprendizado supervisionado de redes neurais, empregam estratégias para garantir que cada classe esteja adequadamente representada no conjunto de treinamento. Dessa forma, buscam mitigar a tendência dos modelos de favorecer classes majoritárias, promovendo maior equidade no aprendizado.

A amostragem estratificada é uma técnica comum utilizada para garantir que cada classe seja representada proporcionalmente no conjunto de treinamento (He & Garcia, 2009). Ela pode melhorar significativamente o desempenho do modelo, especialmente em casos de desequilíbrio extremo. No entanto, esta técnica tem a desvantagem de potencialmente ignorar a estrutura natural dos dados, forçando uma distribuição que pode não refletir a realidade do problema em questão. Além disso, quando não há dados anotados a priori, não é possível garantir que os estratos para a amostragem serão correlacionados com a divisão dos dados entre classes.

Uma outra técnica é o aprendizado sensível ao custo, que atribui pesos diferentes às classes, de modo que os erros nas classes minoritárias tenham um custo maior (Elkan, 2001). Essa ponderação pode ajudar a compensar o desbalanceamento, mas a definição do custo ideal pode ser arbitrária e desafiadora, principalmente sem conhecimento prévio dos dados. Além disso, há o desafio de escolher as métricas corretas para se avaliar modelos nesse cenário, em que uma ou mais classes são ponderadas de forma diferente no treinamento.

Outras abordagens buscam reamostrar o conjunto de dados, seja por subamostragem da classe majoritária ou superamostragem da classe minoritária. No entanto, a subamostragem pode levar à perda de informações importantes, enquanto a superamostragem pode introduzir sobreajuste devido à replicação de exemplos.

Em síntese, a busca pelo balanceamento em dados de treinamento é uma tarefa complexa que pode acarretar prejuízos em termos de complexidade, eficácia e representatividade. Neste trabalho, propomos métodos capazes de operar de maneira eficiente em cenários com poucos recursos — mais especificamente, na ausência de dados anotados —, visando construir um modelo inicial que lide com o desbalanceamento de maneira informada. Na seção 3, detalhamos como nossas abordagens propostas buscam superar as limitações das estratégias atualmente disponíveis.

2.3. Otimização de Dados para Treinamento Eficiente de LLMs

Os grandes modelos de linguagem (LLMs) são treinados em conjuntos de dados massivos, tipicamente desbalanceados, que abrangem múltiplos domínios e estruturas linguísticas variadas. Na fase de pré-treinamento, geralmente auto-supervisionada, o objetivo é capturar uma ampla diversidade linguística a partir de corpora extensos. Posteriormente, etapas de refinamento, como alinhamento supervisionado por instrução e aprendizado por reforço, ajustam o modelo para atender a requisitos específicos. Um desafio central nesses estágios é definir estratégias eficazes de mistura de dados que equilibrem representatividade, eficiência computacional e capacidade de generalização.

Estudos recentes indicam que menores conjuntos de dados, quando altamente diversos, podem superar grandes corpora não estruturados, resultando em ganhos expressivos de desempenho (Qin et al., 2025; Albalak et al., 2024; Ge et al., 2024). Métodos como REGMIX (Liu et al., 2025) e DoReMi (Xie et al., 2023) utilizam modelos procuradores (*proxy models*) para otimizar a seleção de dados, demonstrando que misturas de dados bem curadas permitem melhor convergência e generalização em comparação com abordagens baseadas em amostragem uniforme. Da mesma forma, o conjunto de dados FineWeb (Penedo et al., 2024) evidencia que subconjuntos filtrados e diversificados de grandes corpora frequentemente superam os conjuntos originais em *downstream benchmarks*, demonstrando grande relevância do problema de aprendizado com dados balanceados.

Embora este trabalho não se concentre especificamente no treinamento de LLMs, ele se alinha a essa tendência ao propor estratégias de seleção informada que priorizam diversidade e representatividade em detrimento do volume bruto de dados. Projetadas inicialmente para otimizar a anotação humana e o aprendizado com poucos exemplos, essas técnicas podem ser estendidas ao treinamento de LLMs, promovendo maior eficiência na composição das misturas de dados. Ao reduzir redundâncias e favorecer amostras mais representativas, tais abordagens têm o potencial de mitigar vieses introduzidos por distribuições desbalanceadas, promovendo maior generalização e robustez. Como resultado, modelos mais eficientes e de alto desempenho podem ser alcançados com um número menor de *tokens* de treinamento, maximizando o uso dos recursos disponíveis.

3. Seleção Informada de Dados

Para abordar o desafio de decidir quais dados anotar, propomos métodos de seleção informada de dados, que podem ser interpretados como algoritmos de ordenação quando executados integralmente. A seleção aleatória de dados frequentemente resulta em distribuições desbalanceadas de classes, causando a sub-representação de certas classes. Os métodos aqui propostos mitigam essa questão, mesmo quando os rótulos não são conhecidos antes do processo de anotação.

Neste trabalho, um conjunto de dados é composto por palavras, frases ou documentos completos que precisam ser anotados e serão referidos como "documentos". A Fig. 1 ilustra a arquitetura proposta para a seleção informada dos dados usados no processo de anotação. O conjunto inicial de documentos consiste em dados não anotados, os quais são selecionados pelos diferentes métodos aqui propostos. Os dados selecionados são então submetidos a um processo de anotação humana e, posteriormente, utilizados no treinamento *few-shot* de um classificador de textos. Cada método de seleção é construído com base em heurísticas distintas, detalhadas nas respectivas seções e resumidas aqui:

- **Busca Semântica Reversa (RSS):** avalia a similaridade semântica e prioriza documentos com baixa similaridade em relação aos já selecionados (Seção 3.1);
- **Clusterização Ordenada (OC):** envolve o agrupamento por densidade de representações vetoriais (*embeddings*) e seleciona sistematicamente um documento de cada grupo com base em seu tamanho (Seção 3.2);
- **Similaridade Lexical Limitada (LLS):** emprega amostragem aleatória para escolher documentos com menor similaridade lexical, excluindo aqueles que compartilham muitos n-gramas comuns (Seção 3.3);
- **Amostragem Estratificada (SS):** emprega agrupamento de representações vetoriais (*embeddings*) utilizando algoritmos baseados em distância e realiza uma amostragem estratificada dos grupos para garantir uma representação proporcional de cada um (Seção 3.4);

A avaliação desses métodos visa identificar uma distribuição mais balanceada entre as classes-alvo, preparando o conjunto de dados para o processo de anotação humana. Nesse contexto, cada classe deve idealmente receber n_{shots} documentos para o processo de anotação. Utilizamos a amostragem aleatória como linha de base para comparação.

A seleção informada de dados é formalmente definida da seguinte forma. Seja $\mathcal{D}$ um conjunto de documentos d_i. Seja E o conjunto de representações vetoriais de cada documento $d_i \in \mathcal{D}$. Seja $C = \{c_1, \ldots, c_{n_{\text{classes}}}\}$ o conjunto de classes-alvo para a tarefa de classificação, enquanto n_{shots} é o número-alvo

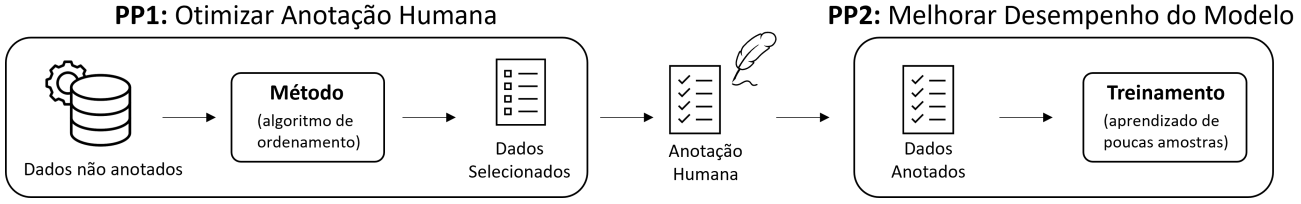

Figura 1: Arquitetura proposta para a seleção informada de dados. O módulo inicial visa otimizar o processo de anotação humana, selecionando os dados e balanceando as classes de interesse. Os dados selecionados são então anotados e utilizados em um processo de treinamento de um classificador *few-shot*. Todo o processo objetiva melhorar o desempenho do classificador.

de documentos anotados por classe-alvo, para treinamento supervisionado do classificador. O conjunto de documentos anotados é

$$D_a = \{D_a^{c_1}, D_a^{c_2}, \ldots, D_a^{c_{n_{\text{classes}}}}\},$$

com $|D_a| = |D_a^{c_1}| + |D_a^{c_2}| + \ldots + |D_a^{c_{n_{\text{classes}}}}|$.

Idealmente, queremos que $|D_a^{c_i}| = n_{\text{shots}}$. Assim, cada método de Seleção Informada de Dados é uma função f:

$$\{D_a^{c_1}, \ldots, D_a^{c_{n_{\text{classes}}}}\} = f(\mathcal{D}, n_{\text{shots}}, n_{\text{classes}}). \quad (1)$$

A **taxa de sobreanotação** θ é definida como o excesso de documentos anotados com o respectivo método utilizado até o número-alvo n_{shots} de documentos anotados para cada classe $c_i \in C$, com $|C| = n_{\text{classes}}$,

$$\theta = |D_a|/(n_{\text{classes}} * n_{\text{shots}}). \quad (2)$$

3.1. Busca Semântica Reversa (RSS)

Dado um conjunto de documentos $d_i \in \mathcal{D}$, seu respectivo conjunto de representações vetoriais $e_i \in E$, e uma função de similaridade por cosseno entre pares de vetores, $sim_{\cos}(e_i, e_j)$, RSS calcula a matriz de similaridade entre todos os pares de vetores de E. A matriz de similaridade S é uma matriz $|\mathcal{D}| \times |\mathcal{D}|$ cujo elemento (i, j) é igual à similaridade $sim_{\cos}(e_i, e_j)$ entre $e_i, e_j \in E$, com e_i e e_j sendo as representações vetoriais de $d_i, d_j \in \mathcal{D}$, respectivamente. RSS inicialmente seleciona os dois documentos com a menor similaridade e coloca ambos em um novo conjunto chamado $\mathcal{D}_{\text{selecionado}}$. Em seguida, iterativamente, RSS continua a selecionar o próximo elemento mais dissimilar do restante do conjunto $\{\mathcal{D} - \mathcal{D}_{\text{selecionado}}\}$. RSS para quando $|\mathcal{D}_{\text{selecionado}}| = |\mathcal{D}|$. Na verdade, RSS ordena os documentos em $\mathcal{D}$ com base na sua dissimilaridade. A ideia é que o processo de anotação seja realizado para cada documento no novo conjunto gerado $\mathcal{D}_{\text{selecionado}}$, em ordem, até que pelo menos n_{shots} sejam obtidos para cada uma das n_{classes}. O pseudocódigo desse método está especificado no Algoritmo 1.

Algoritmo 1 Busca Semântica Reversa (RSS)

Entrada: $\mathcal{D}, E, f, n_{\text{classes}}$
Saída: $\mathcal{D}_{\text{selecionado}}$
criar $\mathcal{D}_{\text{selecionado}} \leftarrow \emptyset$
for each $(e_i, e_{j\neq i}) \in \{E \times E\}$ **do**
  $S_{ij} \leftarrow sim_{\cos}(e_i, e_j)$ ▷ matriz de similaridade
end for
$i, j \leftarrow argmin(S)$ ▷ dois documentos c/ ↓ similaridade
$\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup \{d_i, d_j\}$
$w \leftarrow [i, j]$ ▷ janela de documentos
$soma\_Sw \leftarrow S[i, :] + S[j, :]$ ▷ Soma das linhas de S em w
while $|\mathcal{D}_{\text{selecionado}}| < |\mathcal{D}|$ **do**
  $k \leftarrow argmin(soma\_Sw) \notin \mathcal{D}_{\text{selecionado}}.index$
  $w \leftarrow w \cup [k]$
  $\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup [d_k]$
  $soma\_Sw \leftarrow soma\_Sw + S[k, :]$
  if $|w| > n_{\text{classes}}$ **then**
    $l \leftarrow w.pop(0)$
    $soma\_Sw \leftarrow soma\_Sw - S[l, :]$
  end if
end while
return $\mathcal{D}_{selecionado}$

3.2. Clusterização Ordenada (OC)

Dado um conjunto de documentos $\mathcal{D}$ e seu respectivo conjunto de vetores E, OC aplica um algoritmo de agrupamento hierárquico baseado em densidade. Esse algoritmo atribui a cada documento uma probabilidade de pertencimento a cada grupo, refletindo a chance do documento estar associado a determinado grupo. Em seguida, OC ordena os grupos com base em seus tamanhos, definidos pelo número de documentos que pertencem a cada um.

O processo de seleção opera de forma iterativa: OC identifica, dentro de cada grupo, o documento com a menor probabilidade de pertencimento, começando pelo maior grupo e avançando em ordem decrescente de tamanho. Esse documento é então removido de seu grupo original e inserido, na ordem de remoção, no conjunto $\mathcal{D}_{\text{selecionado}}$. O procedimento continua até que todos os grupos estejam vazios.

A partir do conjunto $\mathcal{D}_{\text{selecionado}}$ gerado, os documentos são anotados sequencialmente até que cada uma das n_{classes} tenha pelo menos n_{shots} documentos anotados. O pseudocódigo detalhado desse método está apresentado no Algoritmo 2.

Algoritmo 2 Clusterização Ordenada (OC)

Entrada: $\mathcal{D}, E, n_{\text{classes}}$
Saída: $\mathcal{D}_{\text{selecionado}}$
criar $\mathcal{D}_{\text{selecionado}} \leftarrow \emptyset$
aplicar algoritmo de agrupamento em E para formar grupos $C = \{C_1, C_2, \ldots, C_{n_{\text{classes}}}\}$
calcular probabilidade de pertencimento para cada $d_i \in \mathcal{D}$ para cada grupo C_j
ordenar grupos C por tamanho, em ordem decrescente
while $|C| > 0$ **do**
  for each grupo $C_j \in C$ **do**
    encontrar $d_i \in C_j$ com a menor probabilidade de pertencimento
    $\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup d_i$
    remover d_i de C_j
    if $|C_j| = 0$ **then**
      remover C_j de C
    end if
  end for
end while
return $\mathcal{D}_{\text{selecionado}}$

3.3. Similaridade Lexical Limitada (LLS)

Dado um conjunto de documentos $\mathcal{D}$, uma função de comparação lexical $g(d_1, d_2)$ (baseada na pontuação BLEU (Papineni et al., 2002), pontuação ROUGE (Lin, 2004) ou outras métricas de pontuação lexical) e um valor de limiar β, LLS escolhe o primeiro documento d_i aleatoriamente e o insere no conjunto inicialmente vazio $\mathcal{D}_{\text{selecionado}}$. LLS então procede escolhendo o próximo documento d_{i+1} aleatoriamente, descartando-o se $g(d_{i+1}, d_i) > \beta$ e mantendo-o caso contrário. LLS para quando não houver mais documentos para selecionar.

Similar aos métodos RSS e OC, o conjunto gerado pode ter muitos elementos. Assim, a anotação ocorre removendo documentos de $\mathcal{D}_{\text{selecionado}}$, na ordem em que foram inseridos em $\mathcal{D}_{\text{selecionado}}$, até que pelo menos n_{shots} sejam obtidos para cada uma das n_{classes}. O pseudocódigo desse método está especificado no Algoritmo 3.

3.4. Amostragem Estratificada (SS)

A Amostragem Estratificada (SS) é um método estatístico que garante que uma amostra reflita com precisão a população, dividindo os dados em subgrupos distintos (estratos) com base em características específicas e, em seguida, realizando amostragem aleatória de cada estrato de maneira proporcional. No contexto do nosso problema, não é trivial adaptar esse método para se tornar um algoritmo de ordenação. Propomos, portanto, adaptar o processo de estratificação como um agrupamento de representações vetoriais (*embeddings*): primeiro, agrupam-se os vetores para formar estratos que capturem as semelhanças semânticas entre documentos. Determina-se então o tamanho de cada grupo e, depois, realiza-se uma amostragem aleatória dentro de cada grupo, proporcional ao seu tamanho no conjunto de dados. Ao contrário do Agrupamento Ordenado (OC), que seleciona documentos com base na probabilidade de pertencimento aos grupos para maximizar a diversidade, SS foca em alcançar uma amostra representativa mantendo a representação proporcional de cada grupo.

Algoritmo 3 Similaridade Lexical Limitada (LLS)

Entrada: $\mathcal{D}$, g, β
Saída: $\mathcal{D}_{selecionado}$
criar $\mathcal{D}_{\text{selecionado}} \leftarrow \emptyset$
$d_i \leftarrow random\_choice(\mathcal{D})$
$\mathcal{D} \leftarrow \mathcal{D} - d_i$
$last \leftarrow d_i$
$\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup d_i$
while $0 < |\mathcal{D}|$ **do**
  $d_i \leftarrow random\_choice(\mathcal{D})$
  $\mathcal{D} \leftarrow \mathcal{D} - d_i$
  if $g(last, d_i) \leq \beta$ **then**
    $\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup d_i$
    $last \leftarrow d_i$
  end if
end while
return $\mathcal{D}_{\text{selecionado}}$

Dado um conjunto de documentos $\mathcal{D}$ e seu respectivo conjunto de vetores E, SS aplica um algoritmo de agrupamento particional onde, de forma heurística, o número de grupos é definido como n_{classes}. Em seguida, SS define a frequência de amostragem de cada grupo, cujo valor é obtido a partir das proporções em tamanho de cada um. A título de exemplo, dados três grupos, um com 60% das amostras, outro com 30% e o último com 10%, serão selecionadas, iterativamente, 6 amostras do primeiro grupo, 3 amostras do segundo e 1 do terceiro, adicionando-as ao conjunto $\mathcal{D}_{\text{selecionado}}$ e removendo-as dos respectivos grupos para evitar seleção duplicada de dados. Define-se como critério de parada a situação onde $\mathcal{D}_{\text{selecionado}}$ obtenha ao menos n_{shots} para cada uma das n_{classes}. O pseudocódigo desse método está especificado no Algoritmo 4.

Algoritmo 4 Amostragem Estratificada (SS)

Entrada: $\mathcal{D}, E, n_{\text{classes}}, n_{\text{shots}}$
Saída: $\mathcal{D}_{\text{selecionado}}$
criar $\mathcal{D}_{\text{selecionado}} \leftarrow \emptyset$
aplicar algoritmo de agrupamento em E para formar grupos $C = \{C_1, C_2, \ldots, C_{n_{\text{classes}}}\}$
calcular as proporções dos grupos com base em seus tamanhos
definir $num\_amostras[1..n_{\text{classes}}] \leftarrow$ menores inteiros que respeitem as proporções dos tamanhos dos grupos
while $|\mathcal{D}_{\text{selecionado}}| < n_{\text{classes}} \times n_{\text{shots}}$ **do**
  for $i = 1$ **até** n_{classes} **do**
    $\mathcal{D}_{\text{selecionado}} \leftarrow \mathcal{D}_{\text{selecionado}} \cup \{num\_amostras[i] \text{ de } C_i\}$
    remover as $num\_amostras[i]$ de C_i
  end for
end while
return $\mathcal{D}_{\text{selecionado}}$

4. Configuração Experimental

Esta seção descreve a configuração experimental para a avaliação da nossa arquitetura proposta de seleção informada de dados. A avaliação é conduzida em cinco conjuntos de dados de classificação de texto, selecionados para explorar diferentes graus de desbalanceamento de dados, diversidade de classes, idioma e domínio.

4.1. Conjuntos de Dados

Utilizamos os seguintes conjuntos de dados em nossos experimentos:

- **AgNews** (Zhang et al., 2015): Um conjunto de dados de notícias com 4 classes e distribuição balanceada de dados. Consiste em 120.000 exemplos de treino e 7.600 exemplos de teste, disponível apenas em inglês.
- **SST5** (Socher et al., 2013): Um conjunto de dados de análise de sentimentos com 5 classes e uma distribuição de dados ligeiramente desbalanceada. Contém 8.544 exemplos de treino, 1.101 exemplos de validação e 2.210 exemplos de teste, disponível em inglês.
- **Emotion** (Saravia et al., 2018): Um conjunto de dados de análise de emoções com 5 classes e distribuição de dados desbalanceada. Inclui 16.000 exemplos de treino e 2.000 exemplos de teste, disponível em inglês.
- **Análise Multilíngue de Sentimentos (MSA)**[2]: Um conjunto de dados multilíngue de análise de sentimentos com 3 classes e distribuição de dados balanceada. Utilizamos o subconjunto em português deste conjunto de dados, que contém 1.839 exemplos de treino e 870 exemplos de teste.
- **BRNews**[3]: Um conjunto de dados de notícias em português brasileiro com 19 classes e distribuição de dados desbalanceada. Compreende 176.114 exemplos de treino e 176.114 exemplos de teste, disponível apenas em português.

As divisões de treino e teste são utilizadas para treinamento e avaliação, a menos que especificado de outra forma. Uma visão geral desses conjuntos de dados é fornecida na Tabela 1. A escolha desses conjuntos de dados visa isolar e examinar variáveis-chave de distribuição de dados. Nosso foco está em examinar o impacto de fatores como o número de amostras por classe, a quantidade de classes dentro de cada conjunto de dados, o grau de desbalanceamento dos dados e o idioma (inglês ou português) nos resultados dos métodos de seleção informada de dados.

Dataset	# docs	classes	Balanceamento	Língua
AgNews	127600	4	balanceado	En
SST5	11855	5	pouco desbalanceado	En
Emotion	18000	6	desbalanceado	En
MSA	3033	3	balanceado	Pt
BRNews	352228	19	muito desbalanceado	Pt

Tabela 1: Resumo das Características dos Conjuntos de Dados utilizados

4.2. Perguntas de Pesquisa

No presente estudo, objetiva-se responder a perguntas de pesquisa específicas por meio de diferentes configurações experimentais, cada uma concebida para fornecer um melhor entendimento sobre a eficácia dos métodos de seleção informada de dados. Essas configurações experimentais são detalhadas a seguir.

PP1: Qual método permite uma anotação humana mais eficiente?

Para abordar esta questão, simulamos um cenário real onde nenhum dado anotado está inicialmente disponível e os anotadores humanos são necessários para anotar os dados. Comparamos diferentes métodos de ordenação projetados para priorizar a anotação e, aproveitando o conhecimento das anotações de referências, quantificamos a taxa de sobreanotação θ (Eq. 2) que cada método pode implicar. Nesse contexto, comparamos o desempenho de

[2]Disponível em: https://huggingface.co/datasets/tyqiangz/multilingual-sentiments

[3]Disponível em: https://huggingface.co/datasets/iara-project/news-articles-ptbr-dataset

nossos métodos de seleção informada de dados com o de uma estratégia de amostragem aleatória, referida como **Random**.

PP2: Qual método resulta em um melhor aprendizado com poucos exemplos?

Para abordar esta segunda questão, voltamos nossa atenção para modelos treinados no conjunto de dados criado no contexto da PP1. O objetivo é determinar se o processo de anotação mais eficiente vem com um custo e pode potencialmente levar a modelos enviesados, resultando em desempenho reduzido em comparação com a amostragem aleatória convencional. Por outro lado, nossa hipótese inicial sugere que a seleção informada de dados, ao aumentar a diversidade de dados, levará à melhoria do modelo, pois fornece mais conhecimento com a mesma quantidade de dados de treinamento.

4.3. Métricas de Avaliação

No contexto da configuração da PP1, a principal métrica de avaliação é a **Taxa de Sobreanotação** (Eq. 2), ou seja, a razão entre a quantidade de dados selecionados e a quantidade efetivamente empregada no treinamento balanceado. Mede o excesso de documentos anotados gerados pelo método até que o alvo desejado n_{shots} seja alcançado para cada classe específica c_i. Esta métrica é relevante, pois em cenários com recursos limitados, busca-se a minimização da anotação excessiva. Para esta métrica, **valores mais baixos** significam mais eficiência.

Quanto à configuração para a PP2, empregamos métricas convencionais comumente usadas em classificação de texto. Estas incluem **Acurácia**, que mede a porcentagem de instâncias classificadas corretamente, e, exclusivamente para o conjunto de dados muito desbalanceado, o **Macro *F1-score***, uma métrica que calcula a média harmônica de precisão e revocação (*recall*) para cada classe e, em seguida, calcula a média desses valores em todas as classes.

4.4. Detalhes de Implementação

Para abordar a PP1, nosso modelo de representação vetorial (*embedding*) escolhido para RSS, OC e SS é o paraphrase-multilingual-mpnet-base-v2[4]. Para realizar o agrupamento em OC, utilizamos o algoritmo HDBSCAN (Campello et al., 2013) e para o agrupamento em SS, utilizamos o KMeans (Arthur & Vassilvitskii, 2007). Utilizamos a pontuação BLEU (Papineni et al., 2002) como função de comparação em LLS. Todo o processo para LLS, SS e Random é executado 10 vezes, de forma idêntica, e os resultados são reportados como valores médios juntamente com intervalos de confiança.

Em relação à PP2, treinamos modelos sob duas configurações distintas para isolar a influência do algoritmo de treinamento para aprendizado de poucos exemplos. Utilizamos a biblioteca *HuggingFace Transformers* (Wolf et al., 2020) e empregamos os seguintes métodos:

- FineTune: Ajustamos o XLM-Roberta-large (Conneau et al., 2019), um modelo de linguagem baseado em codificação (*encoder*) pré-treinada, seguindo procedimentos convencionais de ajuste fino para classificação de sequências. O processo de treinamento abrange 30 épocas com uma taxa de aprendizado de 2×10^{-5}.
- SetFit: Para este método, utilizamos o ajuste fino de *Sentence Transformers* (SetFit) (Tunstall et al., 2022), uma abordagem eficiente para aprendizado de poucos exemplos em modelos baseados em codificação. SetFit gera dinamicamente pares de treinamento a partir dos dados anotados e aproveita a perda contrastiva para treinar o modelo na tarefa de classificação. Como modelo base, também usamos paraphrase-multilingual-mpnet-base-v2.

Os resultados na PP2 para LLS e Random, que exibem comportamento estocástico, são apresentados em termos de valores médios e desvios padrão em todas as 10 execuções. Os experimentos são conduzidos para uma faixa de valores de n_{shots}, especificamente 8, 16, 32 e 64, com um tamanho de lote de 16 para o processo de treinamento.

5. Resultados

Comparamos o desempenho dos nossos métodos propostos de seleção informada de dados com a estratégia de amostragem aleatória nos cinco conjuntos de dados. Os gráficos na Figura 2 mostram os resultados da primeira configuração (PP1), onde medimos a Taxa de Sobreanotação. Os métodos LLS, SS e Random são executados 10 vezes, e os resultados médios (junto com o intervalo de confiança) são apresentados.

5.1. Eficiência na Anotação Humana

Em **conjuntos de dados balanceados** (AgNews e MSA), observamos que **nenhum método supera consistentemente** a linha de base aleatória. Isso pode ser explicado, como mencionamos antes, pela distribuição de classes nesses conjuntos de dados: ambos são fortemente balanceados, o que tende a fa-

[4]Disponível em: https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

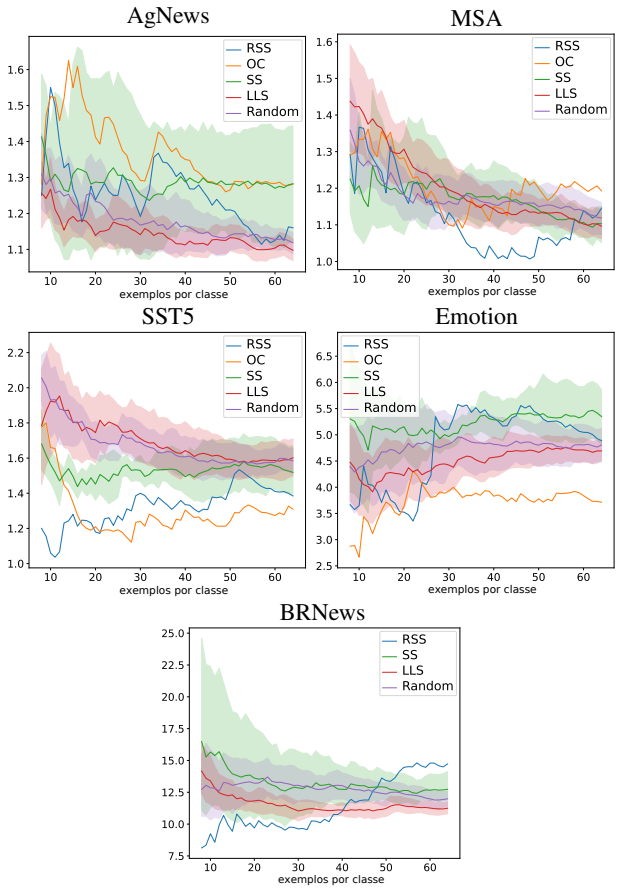

Figura 2: Taxa de sobreanotação em função do número de amostras por classe para diferentes métodos em cada conjunto de dados. As curvas representam a taxa média, enquanto as áreas sombreadas indicam a variabilidade dos resultados nos métodos não determinísticos.

vorecer métodos de amostragem aleatória. É interessante notar que no MSA, quando n_{shots} está na faixa de 30 a 60, RSS seria de fato uma escolha melhor do que a amostragem aleatória (Random). Além disso, nossos métodos são ligeiramente mais competentes no MSA do que no AgNews. O fator idioma pode desempenhar um papel menor aqui: como nosso modelo de *embedding*, embora multilíngue, foi treinado com mais dados em inglês do que em português, suas *embeddings* são menos ajustadas para o idioma português, o que pode explicar porque a busca reversa (RSS) promove variedade por um intervalo maior de n_{shots}, mas eventualmente converge com a maioria dos outros métodos. Além desse possível fator relacionado ao modelo, o idioma não parece ser um fator relevante para nossos métodos de seleção.

Para **distribuições de dados desbalanceadas** (SST5, Emotion), dois de nossos métodos superam consistentemente a amostragem aleatória (Random): Busca Semântica Reversa (RSS) e Agrupamento Ordenado (OC). Observamos uma menor Taxa de Sobreanotação no SST5 e Emotion quando $n_{\text{shots}} < 30$, indicando que tanto RSS quanto OC são mais adequados do que a amostragem aleatória em distribuições desbalanceadas. À medida que aumentamos n_{shots} acima de 30, apenas RSS no conjunto de dados Emotion piora, mas no geral os métodos são mais eficientes na escolha de quais dados anotar, gerando menos excesso de anotações.

Para uma **distribuição fortemente desbalanceada** (BRNews), vemos um comportamento diferente. Observamos que à medida que **o número de classes e o desequilíbrio dos dados aumentam, a Taxa de Sobreanotação aumenta** para todos os métodos testados (BRNews tem 10 vezes mais Sobreanotação do que conjuntos de dados balanceados). Por sua vez, o Agrupamento Ordenado (OC) gera muita Sobreanotação (mais de 6 vezes a linha de base aleatória), e é, portanto, considerado um dado discrepante e excluído do gráfico. Os resultados mostram que **RSS supera consideravelmente a abordagem aleatória** para $n_{\text{shots}} < 40$. Isso se deve, novamente, ao fato de que este conjunto de dados tem um número muito maior de classes, com uma distribuição de documentos por classe muito desbalanceada, muito mais próxima de um cenário real. Nesses cenários, nosso método prospera, gerando até metade das Sobreanotações quando comparado ao método aleatório.

Em todos os casos, observamos que o método Amostragem Estratificada (SS) exibe alta variância, tanto entre diferentes n_{shots} de uma mesma base de dados como entre diferentes bases de dados. Devido a esse resultado e à complexidade de se adaptar o método SS para o cenário deste trabalho, onde as condições iniciais são a ausência de dados anotados e sua progressiva anotação, argumentamos que a adaptação do método de Amostragem Estratificada é inviável para a tarefa em questão, não avaliando este método na Configuração da PP2. Finalmente, como observado também **para todos os conjuntos de dados, nossos métodos e a abordagem aleatória tendem a convergir** quando n_{shots} aumenta além de cerca de 50.

5.2. Desempenho do Modelo

A Figura 3 mostra os resultados da configuração para PP2, onde comparamos o desempenho dos classificadores treinados com dados selecionados pelos métodos da configuração da PP1 (exceto SS). Como OC falha em gerar um excesso factível de anotações para BRNews, ele também é considerado não-viável e, portanto, excluído dos relatórios para essa base de dados.

Como resultado geral, observamos que nossos métodos **OC e LLS falham em superar consistentemente** a amostragem aleatória (Random). No en-

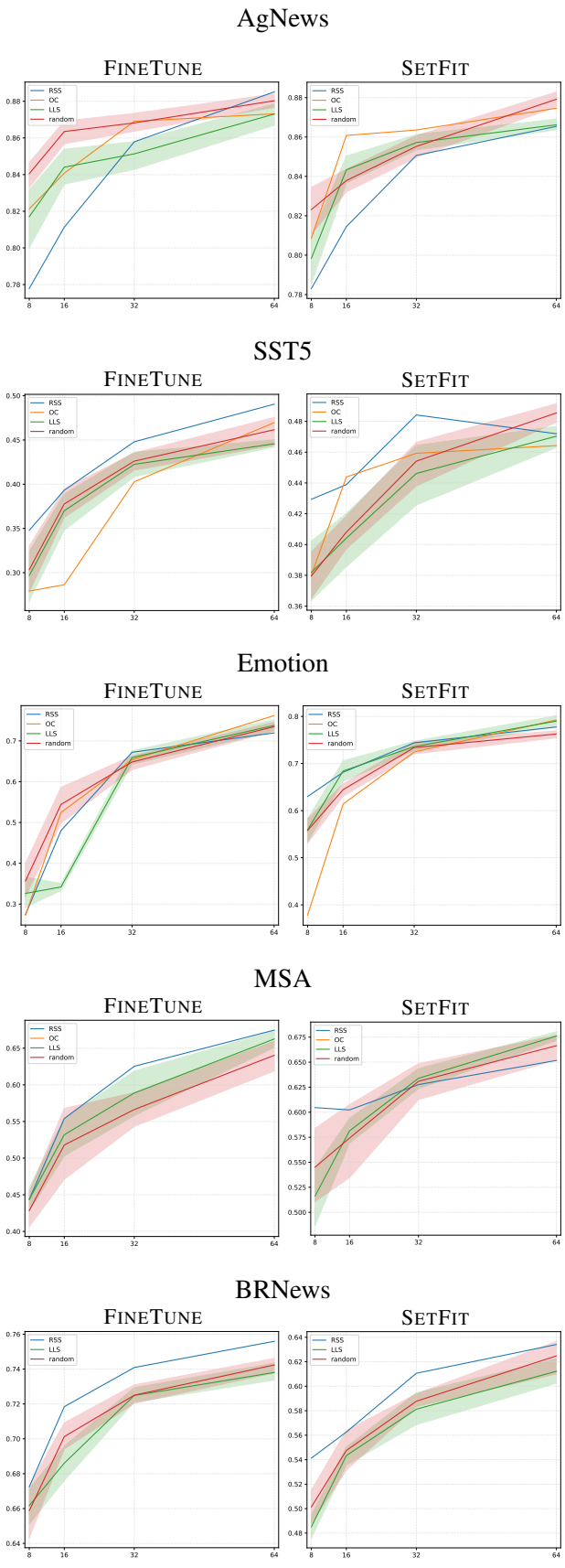

Figura 3: Acurácia dos classificadores treinados com diferentes métodos de seleção de dados (RSS, CC, LLS e Random) utilizando FineTune e SetFit. Os gráficos mostram a variação da acurácia com o aumento do número de exemplos de treinamento por classe nas bases de dados (AgNews, SST5, Emotion, MSA e BRNews).

tanto, **Busca Semântica Reversa (RSS) supera a amostragem aleatória em quase todos os cenários**. Para ambos FineTune e SetFit, RSS é melhor do que a amostragem aleatória para todos os conjuntos de dados, exceto o AgNews, onde a amostragem aleatória apresenta maior acurácia. Uma mistura de muitos fatores pode ser responsável por isso: primeiro, AgNews é balanceado, o que favorece a amostragem aleatória ao selecionar dados de treinamento; segundo, a tarefa de AgNews é simples quando comparada a outros conjuntos de dados, porque suas classes têm características distintas (ou seja, referem-se a temas distintos, como Esportes, Tecnologia, etc.), o que pode ajudar nas fronteiras de decisão do modelo. O outro conjunto de dados balanceado, MSA, não possui essas características distintas para suas classes, as quais, em vez disso, expressam um tipo de graduação (ou seja, Positivo, Neutro, Negativo). Em outras palavras, a tarefa de classificação no MSA é mais difícil, o que significa que selecionar dados com mais variabilidade pode efetivamente aumentar o desempenho do modelo.

Observamos que **quanto maior o grau de desbalanceamento dos dados, mais consistentemente RSS supera a amostragem aleatória**. No entanto, relatar apenas a acurácia em um conjunto de dados fortemente desbalanceado é insuficiente para representar adequadamente o desempenho de um classificador. Assim, a Tabela 2 mostra os resultados do **Macro-F1 Score** para ambos os métodos de treinamento no BRNews. Vemos que para FineTune, **RSS apresenta um desempenho significativamente melhor**, enquanto para SetFit, RSS também mostra uma ligeira melhora quando comparado ao método aleatório, ficando acima do intervalo de confiança apenas para $n_{\text{shots}} = 8$. Isso é um indicativo de que **tanto RSS quanto os métodos aleatórios desempenham quase igualmente bem entre as classes**, desconsiderando o desequilíbrio entre elas. Isso indica que ambos os métodos têm sucesso em selecionar dados diversos para o treinamento do modelo. Ainda assim, **RSS proporciona maior acurácia**, tornando-o o **método recomendado para configurações com poucos recursos**.

Outro resultado importante é **a convergência de todos os métodos quando n_{shots} aumenta**. Como nossos métodos são adequados para a construção de uma primeira versão de um conjunto de dados para Aprendizado Ativo, tanto a Taxa de Sobreanotação quanto o desempenho do modelo convergem quando $n_{\text{shots}} > 64$. Uma razão é que, à medida que o número de dados selecionados aumenta, a diversidade também aumentará. Embora os resultados mostrem que nossos métodos de seleção promovem mais diversidade para menores n_{shots}, qualquer método de seleção que não aplique superamostragem trará diver-

Treinamento	n_{shots}	**RSS**	**Random**
FINETUNE	8	58.6	56.7 ± 2.3
FINETUNE	16	62.0	60.6 ± 0.8
FINETUNE	32	65.2	62.8 ± 0.9
FINETUNE	64	66.8	63.6 ± 0.5
SETFIT	8	46.83	45.0 ± 1.7
SETFIT	16	48.8	48.8 ± 1.9
SETFIT	32	52.3	52.2 ± 1.0
SETFIT	64	55.9	55.6 ± 1.2

Tabela 2: Resultados de Macro-F1 Score em função do número de dados por classe (n_{shots}) e dos métodos RSS e Amostragem Aleatória (Random) para treinamentos com FINETUNE e SETFIT.

sidade se n_{shots} continuar aumentando. Assim, outros métodos superam os nossos no quesito diversidade quando deixamos o domínio de poucos exemplos, ou seja, quando anotamos muitos dados. Isso significa que **quando o interesse é anotar muitos dados por classe, a maioria dos métodos avaliados neste trabalho não são adequados para a seleção**, sendo a amostragem aleatória uma estratégia melhor.

6. Conclusão

Este trabalho propôs uma arquitetura automática de Seleção Informada de Dados que visa selecionar quais dados devem ser anotados por um humano para construir um primeiro conjunto de dados. Simulamos dois cenários, e os resultados experimentais que relatamos mostram que nossa arquitetura é uma opção melhor do que métodos de amostragem aleatória para aprendizado de poucos exemplos. Demonstramos que quanto maior o desbalanceamento no conjunto de dados, mais competente é nosso método, tanto em gerar menos excesso de anotações quanto em melhorar o desempenho do modelo.

Em particular, o método de Busca Semântica Reversa (RSS) mostrou ser o mais competente em experimentos em diferentes idiomas, números e desequilíbrios entre as classes. No entanto, algumas limitações devem ser notadas: primeiro, para a Semelhança Lexical Limitada (LLS), um limite numérico é especificado. Este trabalho não ajustou esse hiperparâmetro, em vez disso, ele foi escolhido inspecionando manualmente os resultados com diferentes limites. Os resultados indicam que ajustar este limite pode melhorar tanto a sobreanotação quanto a acurácia do LLS, pois a variância exibida por esse método em muitos conjuntos de dados é menor do que a do método aleatório. Segundo, usar outra função de comparação, como a pontuação ROUGE, pode ser benéfico. Esse efeito pode ser amplificado em domínios muito especializados, com vocabulário incomum, mas mais experimentos são necessários para confirmá-lo.

Também observamos que os dois métodos que utilizam agrupamento não forneceram resultados consistentes em muitos conjuntos de dados. Como nossos métodos dependem de escolher um documento de cada grupo, quando o número de grupos identificados é alto, os métodos falham em selecionar dados de qualidade. Isso pode ser resolvido combinando agrupamento com RSS ou LLS, e é uma direção para trabalhos futuros. No caso específico da Amostragem Estratificada (SS), a instabilidade do método é indesejada no cenário deste estudo, onde há uma busca por eficiência na anotação. Notamos, entretanto, que nossa implementação é uma adaptação do método e que outras heurísticas de seleção, para este mesmo método, podem ser mais eficazes.

Outra conclusão aponta para um aspecto comum, embora frequentemente negligenciado, do aprendizado de poucos exemplos em cenários com escassez de anotações. A maioria dos trabalhos de poucos exemplos geralmente avalia seus métodos em uma variedade de conjuntos de dados que são principalmente balanceados ou ligeiramente desbalanceados. Existem algumas exceções, mas trabalhos que avaliam em conjuntos de dados desbalanceados não abordam suficientemente as consequências do desbalanceamento. É um fato que, em cenários da vida real, distribuições de dados balanceadas são muito raras. Portanto, argumentamos que trabalhos que lidam com técnicas de aprendizado com poucos exemplos devem considerar esses fatos, relatando métricas conscientes do desbalanceamento, sempre que possível.

Nosso comentário final destaca que, mesmo quando o desbalanceamento entre as classes na população é pequeno, a disponibilidade de dados anotados tende a ser fortemente desigual. Modelos tradicionais de classificação frequentemente assumem um equilíbrio entre as classes nos dados rotulados — uma questão central abordada neste trabalho.

Por fim, este trabalho focou em estender as técnicas propostas por Alcoforado et al. (2024), mas os princípios de seleção baseada em diversidade apresentados neste trabalho podem ser explorados em contextos mais amplos, como a curadoria de dados para o treinamento de LLMs. Os métodos aqui explorados, originalmente projetados para otimizar a anotação em cenários de poucos dados, podem ser utilizados como heurísticas para a composição de misturas de dados em larga escala, abordando um dos principais desafios do pré-treinamento de LLMs: a seleção eficiente de dados diversos e representativos sem necessidade de grandes volumes de amostras. Estudos recentes demonstram que conjuntos de dados menores, mas mais diversificados, frequente-

mente superam corpora massivos e não filtrados em tarefas de generalização e alinhamento de modelos (Qin et al., 2025; Penedo et al., 2024; Liu et al., 2025; Xie et al., 2023; Ge et al., 2024). Assim, uma linha promissora de pesquisa futura seria a incorporação de nossas técnicas de seleção informada no treinamento de LLMs, visando à construção de misturas de dados mais eficientes, tanto na fase de pré-treinamento quanto nas etapas de alinhamento por instrução supervisionada ou aprendizado por reforço.

Limitações

Uma limitação deste trabalho é a predominância de conjuntos de dados em inglês e português brasileiro nos experimentos. Embora os métodos propostos priorizem a diversidade semântica, eles não foram validados em corpora que integrem diferentes variantes do português, como o português europeu, nem diversas línguas. Além disso, este estudo não realizou uma avaliação comparativa entre diferentes modelos de *embeddings*. Os métodos aqui utilizados dependem de modelos de *embeddings* multilíngues, que podem ser menos ajustados ao português em relação ao inglês. Trabalhos contemporâneos como Muennighoff et al. (2023), Lee et al. (2025) e Duquenne et al. (2023) comparam diferentes modelos de *embeddings* com base em seu desempenho em tarefas de classificação, agrupamento e em cenários multilíngues. Trabalhos futuros poderiam investigar o desempenho dos métodos em corpora mistos, explorando variações na estrutura linguística e semântica entre as variantes da língua, bem como investigar modelos mais adequados para tarefas de agrupamento e classificação multilíngue.

Por fim, alguns métodos testados exigem a definição de hiperparâmetros, particularmente a Similaridade Lexical Limitada (LLS), que requer um limiar de similaridade definido pelo usuário. Nos experimentos, o ajuste desse parâmetro foi feito manualmente, sem uma busca sistemática, o que consiste em um trabalho futuro.

Agradecimentos

Este trabalho é parcialmente apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) do Brasil (Processo 312360/2023-1), pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Código de Financiamento 001), e pelo *Center for Artificial Intelligence* USP-IBM-FAPESP (Projeto FAPESP 2019/07665-4), Brasil. Este trabalho também é financiado por Fundos Nacionais através da agência de financiamento portuguesa, FCT - Fundação para a Ciência e a Tecnologia, no âmbito do projeto 10.54499/LA/P/0063/2020. Israel Campos Fama é financiado pela Secretaria da Fazenda do Estado do Rio Grande do Sul (Sefaz-RS). Bárbara Dias Bueno é financiada pela bolsa de pesquisa de graduação do Programa Unificado de Bolsas (PUB) da Universidade de São Paulo (Projeto PUB 2117/2023).

Referências

Albalak, Alon, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto & William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*

Alcoforado, Alexandre, Thomas Palmeira Ferraz, Rodrigo Gerber, Enzo Bustos, André Seidel Oliveira, Bruno Miguel Veloso, Fabio Levy Siqueira & Anna Helena Reali Costa. 2022. ZeroBERTo: Leveraging zero-shot text classification by topic modeling. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 125–136. 10.1007/978-3-030-98305-5˙12

Alcoforado, Alexandre, Thomas Palmeira Ferraz, Lucas Hideki Okamura, Israel Campos Fama, Arnold Moya Lavado, Bárbara Dias Bueno, Bruno Veloso & Anna Helena Reali Costa. 2024. From random to informed data selection: A diversity-based approach to optimize human annotation and few-shot learning. Em *16th International Conference on Computational Processing of Portuguese (PROPOR)*, 492–502.

Arthur, David & Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. Em *18th Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.

Basile, Angelo, Guillermo Pérez-Torró & Marc Franco-Salvador. 2021. Probabilistic Ensembles of Zero- and Few-Shot Learning Models for Emotion Classification. Em Ruslan Mitkov & Galia Angelova (eds.), *International Conference on Recent Advances in Natural Language Processing (RANLP)*, 128–137.

Beijbom, Oscar. 2014. Random sampling in an age of automation: Minimizing expenditures through balanced collection and annotation. ArXiv [cs.CY/cs.LG/stat.ME]. 10.48550/arXiv.1410.7074

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon

Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. Em *34th Conference on Neural Information Processing Systems*, 1877–1901.

Campello, Ricardo J. G. B., Davoud Moulavi & Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. Em *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 160–172. 10.1007/978-3-642-37456-2˙14

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. 10.48550/arXiv.1911.02116

Ducoffe, Melanie & Frederic Precioso. 2018. Adversarial active learning for deep networks: a margin based approach. ArXiv [cs.LG/cs.CV/stat.ML]. 10.48550/arXiv.1802.09841

Duquenne, Paul-Ambroise, Holger Schwenk & Benoît Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations. ArXiv [cs.CL]. 10.48550/arXiv.2308.11466

Elkan, Charles. 2001. The foundations of cost-sensitive learning. Em *International Joint Conference on Artificial Intelligence (IJCAI)*, 973–978.

Ferraz, Thomas Palmeira, Alexandre Alcoforado, Enzo Bustos, André Seidel Oliveira, Rodrigo Gerber, Naíde Müller, André Corrêa d'Almeida, Bruno Miguel Veloso & Anna Helena Reali Costa. 2021. DEBACER: a method for slicing moderated debates. Em *XVIII Encontro Nacional de Inteligência Artificial e Computacional*, 667–678. 10.5753/eniac.2021.18293

Ferraz, Thomas Palmeira, Marcely Zanon Boito, Caroline Brun & Vassilina Nikoulina. 2024a. Multilingual distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. Em *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10716–10720. 10.1109/ICASSP48485.2024.10447520

Ferraz, Thomas Palmeira, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal & Nanyun Peng. 2024b. LLM self-correction with DECRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. Em *Findings of the Association for Computational Linguistics (EMNLP)*, 7773–7812. 10.18653/v1/2024.findings-emnlp.458

Ge, Yuan, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing Chen, Hao Yang, Bei Li, Tong Xiao & JingBo Zhu. 2024. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 464–478. 10.18653/v1/2024.emnlp-main.28

He, Haibo & Edwardo A. Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9). 1263–1284. 10.1109/TKDE.2008.239

Hovy, Dirk, Taylor Berg-Kirkpatrick, Ashish Vaswani & Eduard Hovy. 2013. Learning whom to trust with MACE. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 1120–1130.

Hsueh, Pei-Yun, Prem Melville & Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. Em *North American Chapter of the Association for Computational Linguistics (NAACL)*, 27–35.

Karpinska, Marzena, Nader Akoury & Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1265–1285. 10.18653/v1/2021.emnlp-main.97

Kee, Seho, Enrique del Castillo & George Runger. 2018. Query-by-committee improvement with diversity and density in batch active learning. *Information Sciences* 454–455. 401–418. 10.1016/j.ins.2018.05.014

Lee, Chankyu, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro & Wei Ping. 2025. NV-embed: Improved techniques for training LLMs as generalist embedding models. Em *13th International Conference on Learning Representations*,

Li, Xirong, Efstratios Gavves, Cees G. M. Snoek, Marcel Worring & Arnold W. M. Smeulders. 2011. Personalizing automated image annotation using cross-entropy. Em *ACM International Conference on Multimedia*, 233–242. 10.1145/2072298.2072330

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. Em *Text Summarization Branches Out*, 74–81.

Liu, Qian, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang & Min Lin. 2025. Regmix: Data mixture as regression for language model pre-training. Em *The Thirteenth International Conference on Learning Representations*,

Muennighoff, Niklas, Nouamane Tazi, Loic Magne & Nils Reimers. 2023. MTEB: massive text embedding benchmark. Em *17th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2014–2037. 10.18653/v1/2023.eacl-main.148

Nowak, Stefanie & Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. Em *International Conference on Multimedia Information Retrieval*, 557–566. 10.1145/1743384.1743478

Papineni, Kishore, Salim Roukos, Todd Ward & Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. Em *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318. 10.3115/1073083.1073135

Penedo, Guilherme, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra & Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. Em *Advances in Neural Information Processing Systems (NeurIPS)*, 30811–30849.

Qin, Yulei, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li & Xing Sun. 2025. Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. *Transactions on Machine Learning Research*

Ren, Pengzhen, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen & Xin Wang. 2021. A survey of deep active learning. *ACM Computing Surveys* 54(9). 10.1145/3472291

Saravia, Elvis, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu & Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3687–3697. 10.18653/v1/D18-1404

Sener, Ozan & Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. Em *International Conference on Learning Representations (ICLR)*,

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng & Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1631–1642.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale et al. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv [cs.CL/cs.AI]. 10.48550/arXiv.2307.09288

Tunstall, Lewis, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat & Oren Pereg. 2022. Efficient few-shot learning without prompts. ArXiv [cs.CL]. 10.48550/arXiv.2209.11055

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Xie, Sang Michael, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma & Adams Wei Yu. 2023. DoReMi: Optimizing data mixtures speeds up language model pretraining. Em *Advances in Neural Information Processing Systems (NeurIPS)*, 69798–69818.

Xu, Hanzi, Muhao Chen, Lifu Huang, Slobodan Vucetic & Wenpeng Yin. 2024. X-shot: A unified system to handle frequent, few-shot and zero-shot learning simultaneously in classification. Em *Findings of the Association for Computational Linguistics*, 4652–4665. 10.18653/v1/2024.findings-acl.276

Yang, Jingfeng, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin & Xia Hu. 2023. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. ArXiv [cs.CL/cs.AI/cs.LG]. 10.48550/arXiv.2304.13712

Zhang, Dequan, Pengfei Zhou, Chen Jiang, Meide Yang, Xu Han & Qing Li. 2021. A stochastic process discretization method combing active learning kriging model for efficient time-variant

reliability analysis. *Computer Methods in Applied Mechanics and Engineering* 384. 113990. 10.1016/j.cma.2021.113990

Zhang, Lining, Simon Mille, Yufang Hou, Daniel Deutsch, Elizabeth Clark, Yixin Liu, Saad Mahamood, Sebastian Gehrmann, Miruna Clinciu, Khyathi Raghavi Chandu & João Sedoc. 2023a. A needle in a haystack: An analysis of high-agreement workers on MTurk for summarization. Em *61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 14944–14982. 10.18653/v1/2023.acl-long.835

Zhang, Xiang, Junbo Zhao & Yann LeCun. 2015. Character-level convolutional networks for text classification. Em *Advances in Neural Information Processing Systems (NeurIPS)*,

Zhang, Yating, Yexiang Wang, Fei Cheng, Sadao Kurohashi et al. 2023b. Reformulating domain adaptation of large language models as adapt-retrieve-revise. ArXiv [cs.CL]. 10.48550/arXiv.2310.03328

Zhu, Jingbo, Huizhen Wang, Benjamin K. Tsou & Matthew Ma. 2010. Active learning with sampling by uncertainty and density for data annotations. *IEEE Transactions on Audio, Speech, and Language Processing* 18(6). 1323–1331. 10.1109/TASL.2009.2033421