Aprendizado de máquina para a sugestão automática de correções em anotações de dados textuais

Tatiana F. Pereira
Dept. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brasil
tatiana.franco@aluno.unb.br

Vinícius R. P. Borges
Dept. de Ciência da Computação
Universidade de Brasília
Brasília, DF, Brasil
viniciusrpb@unb.br

***Abstract*—O treinamento de modelos de aprendizado de máquina supervisionado ou semi-supervisionado requer dados rotulados, os quais nem sempre estão disponíveis. Em função disso, surge a necessidade de rotular dados manualmente. a construção de um corpus manualmente anotado é extremamente custoso e costuma envolver múltiplos anotadores. A uniformidade dos rótulos é essencial para o sucesso das tarefas que irão utilizá-los e com isso, é necessário incluir uma etapa de revisão nesse processo, a fim de se garantir a qualidade final do conjunto de dados rotulados. Este trabalho propõe a utilização de técnicas de aprendizado de máquina, como o Support Vector Machine (SVM), para a sugestão automática de anotações a serem corrigidas. O objetivo é reduzir o número de documentos que precisam ser inteiramente revisados por humanos, sem afetar a qualidade final do corpus gerado. Adotou-se o coeficiente de concordância kappa apresentado por Cohen a fim de se avaliar a viabilidade da metodologia proposta. Os resultados indicam que o modelo conseguiu identificar com êxito quais correções sugerir ao revisor.**

***Index Terms*—Anotação de Dados Textuais, Aprendizado de Máquina**

I. Introdução

O treinamento de modelos de aprendizado de máquina supervisionado ou semi-supervisionado requer dados rotulados, os quais nem sempre estão disponíveis. Em função disso, surge a necessidade de rotular dados. A literatura relata que a anotação manual é a abordagem mais apropriada [19]. No entanto, a construção de um corpus manualmente anotado é extremamente custoso, tanto em horas de anotação quanto em questões financeiras [16]. Além disso, o processo de anotação costuma envolver múltiplos anotadores e a uniformidade dos rótulos é essencial para o sucesso das tarefas que irão utilizá-los. Com isso, é necessário que se inclua uma etapa de revisão nesse processo, a fim de se garantir a qualidade final do conjunto de dados rotulados.

Este trabalho propõe a utilização de técnicas de aprendizado de máquina, como o Support Vector Machine (SVM), para a sugestão automática de anotações a serem corrigidas. A hipótese é de que isso reduziria significativamente o número de documentos que precisam ser inteiramente revisados por humanos, sem afetar a qualidade final do corpus gerado. Com isso, o tempo gasto com a revisão de anotações também seria reduzido, tornando o processo de anotação menos custoso.

A fim de se avaliar a viabilidade da metodologia proposta, foram realizados experimentos em um conjunto de dados textuais que foram obtidos a partir da anotação manual de documentos do Diário Oficial do Distrito Federal (DODF). Comparou-se o nível de concordância entre a versão final dos dados rotulados com a classificação feita utilizando SVM por meio do coeficiente kappa (k) proposto por Cohen[4]. Também calculou-se a precisão e o recall para as sugestões de correções feitas pelo modelo.

Este artigo está organizado da seguinte forma. A Seção II fornece uma visão geral da revisão de literatura, abordando trabalhos relacionados a máquinas de vetores, a integração entre anotações manuais e semi-automáticas, a qualidade do corpus e detecção de erros de anotação. A Seção III descreve a metodologia proposta para a revisão de anotações de dados textuais com sugestões automáticas de correções fornecidas pelo modelo SVM. Em seguida, a descrição dos experimentos de validação e análise dos resultados obtidos são apresentadas na Seção IV. Por fim, a conclusão e trabalhos futuros podem ser encontradas na Seção V.

II. Trabalhos Relacionados

A. Integração entre anotações manuais e semi-automáticas

Dados anotados manualmente servem de base para uma quantidade considerável de tarefas de processamento de linguagem natural. Todavia, sua obtenção pode ultrapassar o orçamento de projetos de pesquisa, sendo necessário buscar meios de tornar essa tarefa menos custosa[6]. Uma alternativa é integrar anotações manuais com técnicas de aprendizado de máquina.

Brum et al. [2] apresentam uma estrutura semi-supervisionada para o processo de anotação de um corpus de análise de sentimentos. O método proposto utiliza uma pequena porção de dados anotados e estendendo-os com um grande conjunto de documentos não rotulados, reduzindo o esforço humano na anotação e fornecendo um meio-termo entre a anotação manual e automática de um grande conjunto de dados. Para fins de avaliação, treinou-se vários classificadores binários com o corpus gerado e os resultados mostram que em alguns casos foi possível preservar a qualidade da anotação do corpus original no corpus resultante.

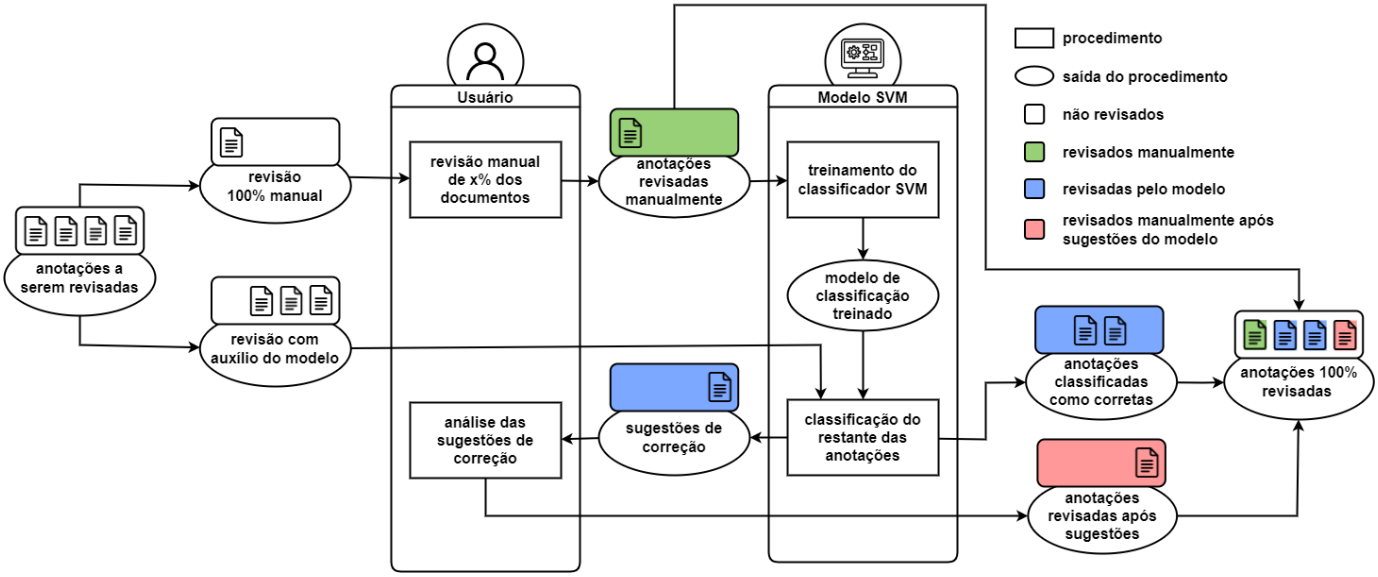

Fig. 1: Fluxograma do método proposto.

Haunss et al.[9] investigam a integração do aprendizado de máquina no fluxo de trabalho de anotação de discursos políticos com o intuito de automatizar parcialmente a anotação e análise de grandes conjuntos de texto. Apesar de os resultados indicarem apenas pequenos aumentos em termos de velocidade, os autores encontraram um aumento moderado na qualidade dos rótulos. Por fim, concluiu-se que uma anotação baseada em inteligência artificial pode fornecer informações confiáveis sobre as redes de discurso com muito menos intervenção humana do que se comparado com a abordagem tradicionalmente manual.

Weeber et al. [18] combinam uma abordagem de aprendizado ativo (AL) com um modelo de linguagem pré-treinado para identificar de maneira semiautomática as categorias de anotação presentes nos documentos de texto fornecidos. A abordagem proposta foi avaliada na tarefa de identificar frames em reportagens e resultados preliminares mostram que o emprego de AL reduz substancialmente o número de anotações manuais necessárias para classificação correta dos dados fornecidos.

Nota-se que os resultados indicam ser possível reduzir o custo das anotações manuais e continuar obtendo informações de qualidade. Um dos objetivos deste trabalho é verificar se o mesmo se aplica a conjunto de dados oriundos de Diários Oficiais, como o Diário Oficial do Distrito Federal.

B. Controle de qualidade do corpus

Existem inúmeras ferramentas de anotação disponíveis, mas dentre seus principais pontos fracos tem-se a ausência de recursos que auxiliem o gerenciamento de dados visando o controle de qualidade. Como a qualidade dos dados de treinamento tem um impacto significativo nos resultados de modelos de processamento de linguagem natural, o controle de qualidade das anotações é essencial. [8, 7]

Tendo isso em vista, há trabalhos na literatura que apresentam ferramentas desenvolvidas para auxiliar o controle de qualidade de conjuntos de dados. Gupta et al. [8] desenvolveram o Data Quality Toolkit, uma biblioteca onde há a integração de algumas métricas de qualidade e técnicas de correção para analisar e aprimorar a prontidão de conjuntos de dados de treinamento estruturados para projetos de aprendizado de máquina.

Por sua vez, Grosman et al. [7] apresentam o ERAS, uma ferramenta de anotação de texto desenvolvida para facilitar e gerenciar o processo de anotação de texto. Além dos principais recursos dos atuais sistemas de anotação convencionais, também foram incluídos outros recursos necessários para melhorar o processo de curadoria das anotações.

C. Detecção de erros de anotação em dados textuais

Em alguns casos, a detecção de erros de anotação é similar à tarefa de deteção de valores atípicos, a qual é particularmente desafiadora em domínios como o de dados textuais. Isso se deve à sua natureza esparsa e de alta dimensão, na qual apenas uma pequena fração das palavras assume valores diferentes de zero. Além disso, muitas palavras em um documento podem ser topicamente irrelevantes para o contexto do mesmo, o que aumenta o ruído nos cálculos de distância. [10]

Ao longo dos últimos anos, foram realizados diversos estudos voltados para aplicabilidade de inteligência artificial em tarefas que visam auxiliar na detecção de erros de anotação. Popović et al. [14] relatam os benefícios de juntar a classificação automática e manual de erros a fim de facilitar a complexa tarefa de anotação manual de erros, bem como o desenvolvimento de ferramentas automáticas que visam sua classificação. Além disso, também é apresentado um corpus contendo textos de idiomas de diferentes origens e domínios, juntamente com suas traduções geradas automaticamente em

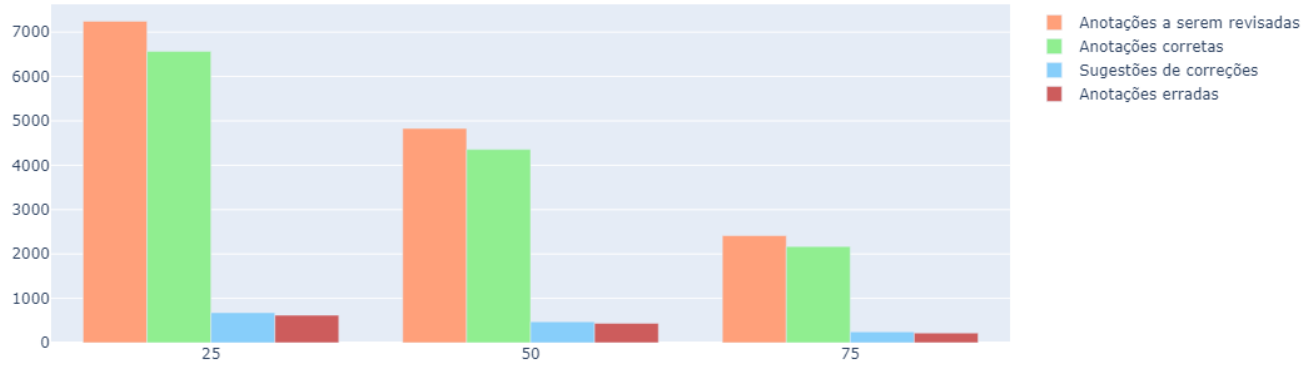

Fig. 2: Comparação da quantidade de anotações que seria preciso revisar.

vários idiomas, suas versões pós-editadas e anotações referentes aos erros encontrados nas operações de pós-edição.

Rehbein et al.[15] apresentaram um método para detecção de erros em textos anotados automaticamente que combina um modelo generativo não supervisionado com supervisão humana de aprendizado ativo. Os autores obtiveram resultados que mostram que a metodologia proposta é capaz de detectar erros de anotação com alta precisão e alto recall. Lu et al. [12] também obtiveram êxito ao investigar o uso de máquinas de vetores para a detecção automática de palavras anotadas incorretamente em corpus de leitura em voz alta de um único locutor. O classificador SVM mostrou um ótimo desempenho alcançando uma medida F1 de quase 88%.

Liu et al. [11] avaliam a viabilidade de se aplicar modelos de transformadores na detecção de erros de rotulação em conjuntos de dados morfológicos baseados em tipos que contêm formas de palavras flexionadas. Para tal, diferentes erros artificiais são inseridos nos dados para se avaliar o modelo apresentado pelos autores. Um procedimento similar foi adotado em nosso trabalho ao adicionar erros sintéticos ao conjunto de dados experimentais a fim de se validar o método proposto. Tal abordagem também foi adotada por Plumb et al. [13] ao apresentarem uma *framework* para gerar dataset sintéticos a serem utilizados em métodos de avaliação de classificadores de imagem.

D. Máquina de Vetores de Suporte (SVM)

A Máquina de Vetores de Suporte, ou Support Vector Machine (SVM) é um modelo de aprendizado de máquina supervisionado cuja classificação dos dados é baseada em encontrar o hiperplano de separação que melhor separa as classes do conjunto. Hiperplano são limites de decisão que classificam o conjunto de dados enquanto maximizam a margem. O objeto de busca do treinamento do classificador é o hiperplano com margem máxima, o qual é chamado de hiperplano ótimo [17].

A princípio, o SVM é um classificador binário. No entanto, ele pode ser utilizado em problemas de classificação multiclasse ao se dividir o problema em múltiplos casos de classificação binária [5]. Uma das técnicas mais utilizadas para tal, e que foi adotada neste trabalho, é a one-vs-rest (um contra o resto, em tradução livre). Ela consiste em construir um modelo de classificação para cada classe e treiná-lo com um conjunto de documentos pertencentes àquela classes (rótulo positivo) e seu complemento (rótulo negativo). Em seguida, para cada documento de teste, se aplica os classificadores separadamente. As classes são atribuídas aos documentos de acordo com a classe que obtiver a melhor pontuação, valor de confiança ou probabilidade. [3]

III. Metodologia

A. Sugestão automática de correções

Ao utilizar um modelo de SVM para sugerir automaticamente sugestões de correções a serem feitas durante a revisão de anotações de dados textuais, ele assume o papel de um anotador e classifica o conjunto de dados tendo como referência um conjunto de documentos previamente revisados manualmente.

Conforme ilustrado na figura 1, inicialmente o conjunto de dados é dividio entre documentos que serão revisados integralmente de forma manual e aquele que fornecerão auxílio de sugestões de correções. A anotação que foram revisadas manualmente pelo usuário são utilizadas no treinamento de um classificador SVM.

Nesta etapa, o modelo aprende os padrões utilizados para rotular os documentos. Tendo isso como base, o modelo utiliza o que aprendeu para classificar o restante dos documentos. Em seguida, a classificação do modelo é comparada com o que foi anotado manualmente e são apresentadas ao revisor os possíveis erros no conjunto de dados, incluindo informações

TABLE I: Descrição do resultado das sugestões do modelo SVM.

	n° de anotações a revisar	n° de erros	n° de sugestões	sugestões 100% corretas	sugestões 50% corretas	sugestões erradas	erros não identificados pelo modelo
25%	7245	620	678	612	7	59	1
50%	4830	438	474	433	5	36	0
75%	2415	221	244	220	1	23	0

sobre o texto anotado e uma sugestão de qual seria o rótulo correto.

O corpus final é formado pelos documentos que foram revisados com o auxílio das sugestões de correções em conjunto com os documentos classificados como corretos pelo modelo e os documentos que foram revisados manualmente no início do processo de revisão.

B. Análise de concordãcia entre anotadores

A confiabilidade de um corpus, assim como sua validade e estabilidade, pode ser medida por meio de coeficientes de concordância [1]. Com o intuito de se comparar o nível de concordância entre a versão final dos dados rotulados e a classificação feita utilizando o modelo, adotou-se o uso do coeficiente de concordância kappa (k) proposto por Cohen[4]. Com tal métrica, é possível medir a frequência com que dois anotadores concordam entre si levando em consideração a probabilidade de tal concordância ocorrer por acaso.

O coeficiente kappa proposto por Cohen é calculado de acordo com a fórmula

$$\frac{Po - Pe}{1 - Pe}$$

Onde *Po* é proporção de vezes em que ambos anotaram o documento com o mesmo rótulo e *Pe* a probabilidade de os dois anotadores escolherem o mesmo rótulo caso os dois o escolhessem de forma aleatória. O valor do coeficiente de concordância de Kappa pode variar de 1, quando ambos anotadores concordaram em 100% dos rótulos, e -1, que significaria que foram escolhidos rótulos distintos para todos os documentos. Um valor igual a 0 ocorre caso a concordância entre os anotadores aconteça na mesma frequência do que se eles estivessem apenas supondo quais são os rótulos aleatoriamente.

IV. Resultados Experimentais

A fim de se avaliar a viabilidade da metodologia proposta, usou-se o coeficiente kappa (k) proposto por Cohen[4] para comparar o nível de concordância entre a versão final dos dados rotulados e a classificação feita utilizando o modelo. Também calculou-se a precisão e o recall para a sugestões de correções feitas pelo modelo.

Os experimentos foram realizados em um conjunto de dados textuais que foram obtidos a partir da anotação manual de documentos do Diário Oficial do Distrito Federal, o qual consiste em 9659 anotações pertencentes a 12 classes distintas. Além disso, foram adicionados 821 erros sintéticos aos rótulos dos dados anotados, ficando com o coeficiente kappa entre dados anotados e dados revisados igual a 0.89.

O modelo foi treinado alterando as porcentagens de dados utilizados para treinamento entre 25%, 50% e 75%. Em uma situação na vida real, essas seriam as anotações revisadas inteiramente por humanos. Para todas as três porcentagens, obteve-se o coeficiente kappa de 0.98, precisão acima de 0.90, e recall de identificação de erros de aproximadamente 100%, como descrito na tabela II.

TABLE II: Métricas de avaliação do modelo SVM.

	Cohen kappa	Precisão	Recall
25%	0.9884	0.90	1.0
50%	0.9894	0.91	1.0
75%	0.9876	0.90	1.0

Conforme o que é apresentado na tabela I, no cenário em que apenas 25% das anotações foram revisadas somente por humanos, restaram 2415 a serem revisadas, das quais 620 estavam erradas. O modelo sugeriu 678 correções com seus possíveis rótulos, dessas, 612 estavam totalmente certas, 7 ele errou apenas a sugestão de rótulo e 59 sugestões eram falsos positivos. Por fim, apenas um erro de anotação não foi identificado pelo classificador.

Nota-se que a quantidade de anotações que não precisam de correção manual é significativa nos três cenários, de 25%, 50% e 75%. A medida em que se revisa uma quantidade maior de documentos inicialmente, restam menos documentos com erros de anotação, o que justifica o decréscimo na quantidade de erros encontrados pelo modelo SVM. Mesmo assim, percebe-se que uma redução no esforço necessário para revisar todos os documentos em todos os casos, e a precisão do modelo não foi afetada. Conclui-se que, com o auxílio do modelos de aprendizado de máquina, é possível obter resultados semelhantes ao se revisar 25% ou 75% dos dados textuais.

V. Conclusão

Neste artigo, propõe-se a utilização de técnicas de aprendizado de máquina, como o Support Vector Machine (SVM), para a sugestão automática de anotações a serem corrigidas. Também é descrito em mais detalhes como a metodologia proposta pode ser implementada.

Os resultados mostram que o modelo conseguiu identificar com êxito quais correções sugerir ao revisor. Em menos de

uma hora o método proposto revisou satisfatoriamente um conjunto de dados que se revisado apenas por humanos, poderia levar meses. Isso indica que com o auxílio de aprendizado de máquina é possível reduzir significativamente o número de documentos que precisam ser revisados manualmente, sem afetar a qualidade final do corpus.

Próximos passos incluem experimentos com outros conjuntos de dados, assim como o desenvolvimento de uma interface para a comunicação entre o modelo e o revisor e a integração com visualizações baseadas no posicionamento de pontos.

References

[1] Ron Artstein and Massimo Poesio. "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004.

[2] Henrico Bertini Brum and Maria das Graças Volpe Nunes. "Semi-supervised Sentiment Annotation of Large Corpora". In: *PROPOR*. 2018.

[3] Prabhakar RaghavanHinrich Schütze Christopher D. Manning. *Introduction to Information Retrieval*. Ed. by Cambridge University Press. 2008.

[4] Jacob Cohen. "A Coefficient of Agreement for Nominal Scales". In: *Educational and Psychological Measurement* 20.1 (1960), pp. 37–46. DOI: 10.1177/001316446002000104. URL: https://doi.org/10.1177/001316446002000104.

[5] Koby Crammer and Yoram Singer. "On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines". In: *J. Mach. Learn. Res.* 2 (Mar. 2002), pp. 265–292. ISSN: 1532-4435.

[6] Brett Drury et al. "An Open Source Tool for Crowd-Sourcing the Manual Annotation of Texts". In: *Computational Processing of the Portuguese Language*. Ed. by Jorge Baptista et al. Cham: Springer International Publishing, 2014, pp. 268–273. ISBN: 978-3-319-09761-9.

[7] Jonatas S. Grosman et al. "Eras: Improving the quality control in the annotation process for Natural Language Processing tasks". In: *Inf. Syst.* 93 (2020), p. 101553.

[8] Nitin Gupta et al. "Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets". In: *CoRR* abs/2108.05935 (2021). arXiv: 2108.05935. URL: https://arxiv.org/abs/2108.05935.

[9] Sebastian Haunss et al. "Integrating Manual and Automatic Annotation for the Creation of Discourse Network Data Sets". In: *Politics and Governance* 8.2 (2020), pp. 326–339.

[10] Ramakrishnan Kannan et al. "Outlier Detection for Text Data : An Extended Version". In: *ArXiv* abs/1701.01325 (2017).

[11] Ling Liu and Mans Hulden. "Detecting Annotation Errors in Morphological Data with the Transformer". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 166–174. DOI: 10.18653/v1/2022.acl-short.19. URL: https://aclanthology.org/2022.acl-short.19.

[12] Heng Lu et al. "Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier". In: Sept. 2010, pp. 162–165. DOI: 10.21437/Interspeech.2010-75.

[13] Gregory Plumb et al. *Evaluating Systemic Error Detection Methods using Synthetic Images*. 2022. DOI: 10.48550/ARXIV.2207.04104. URL: https://arxiv.org/abs/2207.04104.

[14] Maja Popović and Mihael Arčan. "PE2rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 27–32. URL: https://aclanthology.org/L16-1005.

[15] Ines Rehbein and Josef Ruppenhofer. "Detecting annotation noise in automatically labelled data". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1160–1170. DOI: 10.18653/v1/P17-1107. URL: https://aclanthology.org/P17-1107.

[16] Rion Snow et al. "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, Oct. 2008, pp. 254–263. URL: https://www.aclweb.org/anthology/D08-1027.

[17] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[18] Franziska Weeber et al. *Assisted Text Annotation Using Active Learning to Achieve High Quality with Little Effort*. 2021. DOI: 10.48550/ARXIV.2112.11914. URL: https://arxiv.org/abs/2112.11914.

[19] Lars Wißler et al. "The Gold Standard in Corpus Annotation". In: *IEEE GSC*. 2014.