INTERNATIONAL DOCTORAL
SCHOOL OF THE USC

Marcos
Fernández Pichel

PhD Thesis

Technologies for extracting and analysing the credibility of health-related online content

Santiago de Compostela, 2023

Doctoral Programme in Information Technology Research

ESCOLA DE DOUTORAMENTO
INTERNACIONAL DA USC

TESE DE DOUTORAMENTO

TECHNOLOGIES FOR EXTRACTING AND ANALYSING THE CREDIBILITY OF HEALTH-RELATED ONLINE CONTENT

Autor

Marcos Fernández Pichel

Directores: David E. Losada Carril, Juan C. Pichel Campos

Titor/a: Juan C. Pichel Campos

PROGRAMA DE DOUTORAMENTO EN TECNOLOXÍAS DA INFORMACIÓN

SANTIAGO DE COMPOSTELA 2023

Á miña familia

Veritas numquam perit.

Séneca

AGRADECEMENTOS

Gustaríame agradecer de forma especial á miña familia, piar fundamental neste proceso. A todos os meus amigos e amigas, especialmente os do pobo e os do CiTIUS. Pero sobre todo a Laura, a miña compañeira de vida, por aguantarme nos peores momentos e por todo o cariño. Tampouco me gustaría rematar sen dirixirme a Juan e David, gracias pola vosa supervisión e consellos. Gracias a todos.

This work has received financial support from: i) project RTI2018-093336-B-C21 (FEDER/Ministerio de Ciencia, Innovación y Universidades – Agencia Estatal de Investigación), ii) project PLEC2021-007662 (MCIN/AEI/10.-13039/501100011033, Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Plan de Recuperación, Transformación y Resiliencia, Unión Europea-Next GenerationEU), iii) project PID2022-137061OB-C22 (Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación, Proyectos de Generación de Conocimiento; suppported by the European Regional Development Fund), iv) Consellería de Educación, Universidade e Formación Profesional (accreditation 2019-2022 ED431G-2019/04, ED431C 2022/19) and the European Regional Development Fund, which acknowledges the CiTIUS-Research Center in Intelligent Technologies of the University of Santiago de Compostela as a Research Center of the Galician University System and v) call of the Xunta de Galicia of 2021 to fund doctoral studies.

Resumo

A evolución notable da World Wide Web e a proliferación das redes sociais revolucionaron indubidablemente a maneira en que se accede e se difunde a información, como corroboran os datos extraídos de múltiples estudos. Esta transformación é, en moitos aspectos, vantaxosa, facilitando un rápido acceso a unha diversa gama de contidos. Non obstante, é de suma importancia estar alerta ante os posibles perigos que poden xurdir desta accesibilidade. Os resultados proporcionados por ferramentas de recuperación automática de información a miúdo poden caracterizarse como non fiables, inadecuados ou de baixa calidade. Isto produce situacións de "disinformation" (involuntaria) ou "misinformation" (intencionada). Esta tese de investigación céntrase na primeira clase e o noso principal obxectivo é avanzar nos métodos automáticos de clasificación e procura para apoiar aos/ás usuarios/as finais no seu acceso a contidos fiables relacionados coa saúde.

Investigacións empíricas anteriores acharon que a interacción con desinformación pode levar a tomar decisións erróneas. Demostrouse que os/as usuarios/as que examinaron resultados de procuras que contiñan información parcial ou totalmente incorrecta eran máis susceptibles a tomar decisións incorrectas.

O efecto da desinformación pode ser máis ou menos perigoso dependendo do contexto, pero é particularmente preocupante no caso da información relacionada coa saúde. Se se aplican os tratamentos sen a supervisión médica apropiada, mesmo poderían derivar danos persoais. Isto quedou patente coa pandemia da COVID-19 en 2020, onde a desinformación sobre a enfermidade e os seus tratamentos proliferou, incluíndo rumores e teorías de conspiración. Algúns investigadores monitorizaron as afirmacións sobre a enfermidade e os seus tratamentos, que se publicaron en medios de comunicación en liña nos primeiros meses de 2020, e atoparon que o 82% delas contiñan algún tipo de información falsa.

Ademais, os motores de procura emerxeron como ferramentas prevalentes para obter consellos médicos, e o uso das redes sociais para acceder a información médica está en crecente aumento, como indican distintos estudos. Arredor do 72% dos usuarios de internet estadounidenses empregan a web para obter información sobre un problema de saúde ou un tratamento. Noutra enquisa, amosouse que o 60% dos cidadáns europeos buscan consellos médicos na rede. Desafortunadamente, as plataformas en liña están cheas de bulos médicos, dietas milagrosas infun-

dadas e recomendacións publicadas por individuos non cualificados. A pesar desta proliferación, propuxéronse ata o de agora poucas solucións para tratar automaticamente a desinformación relacionada coa saúde.

Nos últimos anos, coa proliferación dos chamados "Large Language Models" (LLMs), popularizouse un novo sistema de acceso á información máis conversacional. Estes avanzados modelos de linguaxe emerxen como novas ferramentas de referencia para apoiar múltiples tipos de necesidades de información. As persoas poden recorrer aos LLMs para resolver as súas necesidades médicas de información e, polo tanto, precísase poñelos baixo un escrutinio rigoroso.

Á luz de todas estas preocupacións, é crucial que tanto os/as usuarios/as como os desenvolvedores de ferramentas de acceso á información exerzan a vixiancia e o discernimento ao navegar pola vasta paisaxe de contidos en liña. A responsabilidade recae na comunidade científica, nos responsables políticos e nas empresas tecnolóxicas. Precisamos idear estratexias que mitiguen a diseminación de información falsa, particularmente no dominio da saúde. Ao fomentar unha cultura de pensamento crítico, promover a alfabetización dixital e implementar mecanismos robustos de verificación de feitos, pódese asegurar que os beneficios do aumento da accesibilidade á información non se vexan eclipsados polos seus potenciais perigos.

En liña con estas afirmacións, a credibilidade da información en liña foi obxecto de atención, particularmente na última década. Investigacións previas revisaron os principais métodos automáticos para estimar a credibilidade nas redes sociais, centrando a atención no contido de saúde. Algúns equipos, por exemplo, estudaron a adhesión dos consumidores a un sitio específico en base á súa percepción de credibilidade. Outras liñas de investigación seguiron métodos orientados ao/á usuario/a final para entender como se utilizan as páxinas de resultados de motores de búsqueda (SERPs) para determinar a credibilidade. Por exemplo, estudaron a influenza dos *featured snippets* dos motores de busca na percepción da credibilidade dos/as usuarios/as. Estes estudos encontraron que moitos/as usuarios/as xulgan a información nestes fragmentos como máis creíble que o resto dos resultados.

Máis en liña co núcleo principal desta investigación, algúns equipos centráronse en avaliar a credibilidade do contido relacionado coa saúde na web. Por exemplo, unha investigación analizou un corpus sobre tratamentos alternativos do cancro e atopou que case o 90% contiña afirmacións falsas.

Nesta tese, presentamos un estudo exhaustivo que vai dende a avaliación das aproximacións clásicas ao problema da detección de desinformación de saúde ata o estudo do papel dos LLMs emerxentes neste campo. Comenzamos reproducindo un estudo seminal para clasificar páxinas web médicas en termos de fiabilidade, para logo comparar as capacidades dos modelos clásicos e neurais na identificación da desinformación. Despois destes primeiros pasos, propoñemos o noso propio sistema de recuperación en varias fases para a detección da desinformación e comparámolo cunha solución de vangarda. Tamén dedicamos esforzos a poñer en escrutinio a

fiabilidade dos novos LLMs para ofrecer consellos médicos.

Paralelamente, propoñemos un conxunto de ferramentas auxiliares para o procesamento de contidos masivos en liña. Estas ferramentas complementarias representan demostradores específicos orientados a casos de uso particulares ou elementos de preprocesamento que axudan na detección de desinformación.

Durante o desenvolvemento desta tese, participamos en tres edicións da Text Retrieval Conference (TREC), unha prestixiosa campaña de avaliación internacional. En concreto, participamos activamente na tarefa de Desinformación de Saúde. Isto permitiunos probar os nosos métodos baixo condicións reais de competición, obtendo o noso mellor resultado, un terceiro posto, na edición de 2021. O obxectivo principal desta tarefa é desenvolver métodos de recuperación que promovan resultados de recuperación correctos e cribles sobre a desinformación. A edición de 2020 centrouse na desinformación relacionada coa COVID-19 e o SARS-CoV-2. A nosa participación en 2020 foi exploratoria e limitouse a implementar un clasificador que replicamos seguindo un estudo clásico. En 2021 e 2022, probamos o noso propio sistema de recuperación en varias etapas, que ten en conta varios sinais para estimar a desinformación. En 2022, introduciuse un novo desafío, orientado a estimar automáticamente a resposta correcta para preguntas médicas. Participamos nesta nova tarefa con algunhas variantes que consisten en explotar os principais resultados dun motor de busca ou *promptear* un LLM (GPT-3).

Polo tanto, o obxectivo principal desta tese é avanzar en métodos automáticos de clasificación e busca para axudarlle aos/ás usuarios/as finais no seu acceso a información online relacionada coa saúde. En particular, queremos afrontar os seguintes retos:

- O1. Avaliar a utilidade das **abordaxes clásicas da aprendizaxe automática** para o problema da detección da desinformación médica. Máis específicamente, probar a capacidade dos métodos clásicos de clasificación de texto para apoiar predicións de fiabilidade a nivel de documento. Isto incluirá actividades como estudos de replicabilidade de resultados publicados por equipos de investigación reputados, comparar a capacidade preditora de características baseadas no contido, nos enlaces ou noutras características do documento web, e experimentos de xeneralización para entender a aplicabilidade desta tecnoloxía clásica a outras coleccións externas (incluídos datos de campañas de avaliación ben coñecidas).

- O2. Avaliar os **novos modelos de redes de neuronas (arquitectura de Transformers)** para a estimación de fiabilidade a nivel de documento e comparalos con solucións máis clásicas. Para iso, consideraremos diferentes variables obxectivo relevantes, como a fiabilidade e a lexibilidade, e compararemos as variantes en diferentes condicións de datos de entrenamento.

– O3. Probar a combinación e o poder predictivo de **múltiples fontes de evidencia** para o problema da detección de desinformación en saúde. Entender que fontes de evidencia ou sinais poden axudar a determinar a fiabilidade da información relacionada coa saúde en liña e como combinar estas pezas de evidencia. Consideraremos unha ampla gama de características, incluíndo estimacións a nivel de documento e pasaxe, sinais supervisados e non supervisados baseados en tecnoloxía *Deep Learning*, etapas de re-ranking e diferentes formas de fusión. Os sinais varían dende estimadores de relevancia tradicionais, como BM25, a predictores de fiabilidade sofisticados baseados en arquitecturas Transformer emerxentes.

– O4. Avaliar os novos **modelos xenerativos de linguaxe** e probar as súas habilidades para proporcionar **consellos médicos fiábeis**. Co actual cambio de paradigma dende motores de busca clásicos a axentes de IA conversacional avanzados, é crítico e socialmente relevante determinar as capacidades destas tecnoloxías emerxentes para proporcionar consellos médicos correctos. Queremos probar diferentes contextos e determinar a súa influencia na calidade dos resultados. Tamén pretendemos avaliar cualitativamente estes modelos para entender os seus erros e desvantaxes potenciais.

– O5. Desenvolver un conxunto de **ferramentas** complementarias que axuden no **procesamento de datos sociais masivos**. Como parte desta investigación, queremos deseñar e desenvolver novas arquitecturas e ferramentas capaces de procesar grandes cantidades de texto publicado nas redes sociais. Isto permitirá probar e implementar os métodos de investigación desenvolvidos nesta tese en casos de uso relevantes (por exemplo, conducindo a demostracións de detección de desinformación en certos dominios). Isto tamén implica o desenvolvemento de ferramentas de pre-procesamento de PLN apropiadas que den soporte en múltiples tarefas futuras, incluíndo a detección de desinformación.

En canto á metodoloxía empregada, nesta tese seguíronse as pautas típicas dos proxectos de investigación baseados no método científico. Máis concretamente, desenvolveuse un proceso iterativo, en varias fases, para avanzar nos métodos automáticos de clasificación e procura para apoiar aos/ás usuarios/as finais no seu acceso ao contido fiable relacionado coa saúde. En primeiro lugar, tomamos como punto de partida algúns estudos reputados no campo da predición da fiabilidade para contido relacionado coa saúde. A replicabilidade dos experimentos previos é crucial no avance da ciencia e, polo tanto, tomamos moi en serio a replicabilidade dos métodos existentes como elementos centrais na nosa investigación. A continuación, avaliamos criticamente as capacidades dos métodos de aprendizaxe tradicionais e comparamos contra as novas abordaxes neuronais para a detección da desinformación. De feito, un obxectivo fundamental da investigación é desenvolver novas técnicas e demostrar que acadan unha eficacia mellorada en comparación coas súas precedentes.

Tamén propoñemos o noso propio sistema de recuperación en varias etapas para a detección da desinformación. Nesta etapa, avaliamos a inclusión de diferentes fontes de evidencia para identificar a desinformación e a súa importancia no resultado final de clasificación. Para iso, deseñamos un conxunto exhaustivo de probas empíricas, incluíndo experimentos de ablación.

Para a avaliación dos resultados, o campo da recuperación da información (RI) ten metodoloxías claramente definidas. A existencia de coleccións de probas estandarizadas e métricas de avaliación ocupan un posto preminente nos fundamentos do campo da RI. Dende o punto de vista metodolóxico, seguiremos con coidado os estándares de avaliación existentes, recorreremos a bancos de probas validados e compararemos os resultados contra alternativas competitivas.

Enfrontamos unha aplicación obxectivo innovadora onde é crucial poder cuantificar a calidade da información relacionada coa saúde (por exemplo, tratamentos para a COVID-19 accesibles en liña). Isto introduce a necesidade de ampliar o conxunto de ferramentas de métricas de rendemento. Relacionado con isto, seguimos aquí avances recentes no campo, como as medidas de compatibilidade. Isto será complementado con métricas máis estándar de clasificación e procura, como F1, Precisión Media ou NDCG. Ademais, tamén probamos a calidade das nosas solucións a través da participación na prestixiosa conferencia TREC, máis especificamente na súa Health Misinformation Track. A participación activa en campañas de evaluación ben coñecidas é un valioso activo desta tese.

Por outra banda, esta investigación tamén se centra en formas emerxentes de acceder á información. Estudamos en profundidade as capacidades dos novos LLMs para proporcionar consellos médicos correctos en diferentes contextos. Para iso, seguimos tendencias recentes na aprendizaxe en contexto e comparamos varios LLMs cunha ampla gama de necesidades de información de saúde.

En paralelo ás actividades científicas, esta investigación tivo un forte compoñente tecnolóxico. Desenvolvéronse tecnoloxías para permitir a análise masiva de linguaxe natural en medios web (por exemplo, redes sociais ou sitios web convencionais). Acceder e procesar información en liña é un desafío en si mesmo, e ser capaz de facelo en tempo real é un valor engadido para moitas aplicacións. É necesario eliminar todos os documentos que non son relevantes para a nosa tarefa (por exemplo, contido publicado en liña que non está relacionado coa saúde). Ademais, deben desenvolverse solucións flexibles e eficientes para o procesamento masivo de datos web en tempo real. Isto implica a construción dunha nova plataforma de Big Data que executa de forma nativa aplicacións paralelas en linguaxes como Python ou Perl. Relacionado con isto, propóñense algunhas ferramentas de preprocesamento que poden impulsar o rendemento das tarefas a realizar.

Debemos destacar que as **investigacións realizadas no desenvolvemento desta tese deron lugar ás seguintes publicacións**:

– Marcos Fernández-Pichel, David E. Losada, Juan C. Pichel e David Elsweiler. Reliability

Prediction for Health-Related Content: A Replicability Study, Proceedings de the 3rd European Conference on IR Research, ECIR 2021, Evento virtual, marzo 28–abril 1, 2021. GGS Rating: Clase 2.

– Marcos Fernández-Pichel, David E. Losada, Juan C. Pichel e David Elsweiler. Comparing traditional and neural approaches for detecting health-related misinformation, Proceedings de Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Evento Virtual, setembro 21–24, 2021. GGS Rating: Clase 3.

– Marcos Fernández-Pichel, David E. Losada, Juan C. Pichel e David Elsweiler. CiTIUS at the TREC 2020 Health Misinformation Track, Proceedings de Twenty-Ninth Text REtrieval Conference (TREC 2020) levada a cabo online novembro 16–20, 2020. GGS Rating: Clase 2 (ranking válido no periodo 2018-outubro 2021).

– Marcos Fernández-Pichel, David E. Losada, e Juan C. Pichel. A multistage retrieval system for health-related misinformation detection. Engineering Applications of Artificial Intelligence. Vol. 115, páxinas 105211, 2022. Índice de impacto: 8.000 (JCR 2022) - Q1 Ciencias da Computación e Intelixencia Artificial. Índice de impacto SJR: 1.73 (SJR 2022).

– Marcos Fernández-Pichel, Manuel Prada-Corral, David E. Losada, Juan C. Pichel e Pablo Gamallo. CiTIUS at the TREC 2021 Health Misinformation Track, Proceedings de Thirtieth Text REtrieval Conference (TREC 2021) levado a cabo online novembro 15–19, 2021. GGS Rating: Clase 3 (ranking válido actualmente).

– Marcos Fernández-Pichel, Manuel Prada-Corral, David E. Losada, e Juan C. Pichel. CiTIUS at the TREC 2022 Health Misinformation Track, Proceedings of Thirty-First Text REtrieval Conference (TREC 2022) levado a cabo online novembro 15–19, 2022. GGS Rating: Clase 3 (ranking válido actualmente).

– Marcos Fernández-Pichel, David E. Losada, e Juan C. Pichel. Social Minder: a Tool for Social Media Monitoring and its Use for Detecting COVID-19 Misinformation, Proceedings de 2nd Joint Conference of the Information Retrieval Communities in Europe, CIRCLE 2022, Toulouse 4–7 xullo, 2022.

– Marcos Fernández-Pichel, Manuel Prada-Corral, David E. Losada, Juan C. Pichel, e Pablo Gamallo. An unsupervised perplexity-based method for boilerplate removal. Natural Language Engineering. Páxinas 1–18, 2023. Índice de impacto: 2.500 (JCR 2022) - Q1 Lingüística. Índice de impacto SJR: 0.58 (SJR 2022).

Nesta memoria detállanse as contribucións do traballo desenvolvido e os resultados máis relevantes da experimentación levada a cabo.

En concreto, o capítulo 2 enfócase na presentación de algúns primeiros enfoques ao problema. Levamos a cabo un estudo de replicabilidade dun traballo anterior sobre a clasificación da fiabilidade de páxinas web médicas. Confirmamos as tendencias observadas no estudo orixinal obtendo incluso un mellor rendemento. Tamén avaliamos esta tecnoloxía predicitiva contra dous novos conxuntos de validación e as conclusións permaneceron sendo as mesmas. O clasificador froito deste primeiro estudo foi a base para a nosa pariticipación na prestixiosa campaña de avaliación, TREC 2020 Health Misinformation Track. Porén, esta aproximación clásica de clasificación a nivel documental non xeneralizou ben para os datos do TREC.

Neste mesmo capítulo, inclúese unha comparación entre enfoques tradicionais e neuronais para a detección de desinformación en saúde en liña. Concluiuse que os modelos tradicionais, como Naive Bayes, establecen todavía unha aproximación sólida para algunhas tarefas de clasificación.

No capítulo 3, presentamos o sistema central desta investigación: un sistema de recuperación en múltiples etapas para a detección de desinformación. Este sistema foi desenvolvido íntegramente como parte da nosa investigación. Levouse a cabo un detallado estudo de ablación para avaliar a utilidade do sistema para un asunto socialmente relevante: a detección de desinformación online relacionada coa COVID-19. O noso análise estudou a efectividade das múltiples etapas e demostrou que cada unha delas foi valiosa para mellorar a performance global do sistema. Por outra parte, a fusión de diferentes fontes de evidencia derivou en métodos de estimación da desinformación máis eficientes. Porén, non todas as solucións foron iguais de boas, xa que as aproximacións de fusión non supervisadas foron moito mellores que as que requerían de entrenamento explícito. Tamén demostramos que o noso sistema é competitivo no estado do arte. Neste mesmo capítulo, informamos da nosa participación nas tarefas de TREC 2021 e TREC 2022 HM, utilizando como base o sistema previamente mencionado.

No capítulo 4, pretendemos ir un paso máis alá e avaliamos as capacidades dos novos LLMs para fornecer consellos médicos. Con este fin, analizáronse diferentes prompts e experimentamos con aproximacións zero- e few-shot. Concluíuse que o prompt seleccionado inflúe enormemente no rendemento final. Máis específicamente, os prompts que nesgan os LLMs hacia fontes máis reputadas aportaron os mellores resultados. Por outra banda, os prompts e modelos máis simples víronse beneficiados da inclusión de exemplos in-context. Por último, levouse a cabo un análise de erros que demostrou que os LLMs todavía producen algúns erros preocupantes.

Finalmente, o capítulo 5 describe aspectos que xurdiron en paralelo ao desenvolvemento principal desta investigación. Primeiro, preséntase un sistema de Big Data para o procesamento masivo de datos publicados en redes sociais e a súa posterior avaliación para a detección de desinformación. Tamén se describe unha nova tecnoloxía de preprocesamento de texto que

estima a presenza de texto mal formado, unha tarefa crítica para obter un bo rendemento en calquera tarefa secundaria.

Polo tanto, pódese concluír que os obxectivos da tese, presentados na Sección 1.1, cumpríronse con éxito.

Contents

Resumo **ix**

Abstract **1**

1 Introduction **2**
1.1 Objectives . . . 8
1.2 Methodology . . . 9
1.3 Publications . . . 10
1.4 Dissertation structure . . . 17

2 Classic Approaches to the Health Misinformation Detection Problem **18**
2.1 Reliability Prediction for Health-related Content: A Replicability Study . . . 19
2.1.1 Dataset . . . 20
2.1.2 Experimental setup . . . 23
2.1.3 Results . . . 23
2.1.4 Additional Experiments . . . 24
2.1.5 Conclusions of the additional experiments . . . 28
2.1.6 Participation in the TREC 2020 Health Misinformation (HM) Track . . . 28
2.1.7 Final remarks . . . 32
2.2 Comparing Traditional and Neural Approaches for detecting Health-related Misinformation . . . 33
2.2.1 Dataset . . . 33
2.2.2 Experimental Design . . . 34
2.2.3 Experimental Results . . . 36
2.2.4 Final remarks . . . 40

3 A Multistage Retrieval System for Health-related Misinformation Detection **41**
3.1 Background . . . 42
3.1.1 Combining multiple signals . . . 42
3.1.2 Systems for Health-related misinformation detection . . . 42

3.2 Use case: Detecting COVID-19 misinformation . . . 43
3.3 Input signals . . . 43
3.4 Fusion strategies . . . 49
3.5 Experimental setup . . . 50
3.5.1 Total Recall . . . 51
3.5.2 Ad-hoc Retrieval Task . . . 52
3.5.3 Experimental details . . . 54
3.6 Results . . . 56
3.6.1 Relevance-based search methods . . . 56
3.6.2 Reliability estimation at passage-level . . . 57
3.6.3 Score Fusion . . . 59
3.7 Discussion . . . 64
3.8 Participation in the TREC 2021 and TREC 2022 Health Misinformation (HM) Tracks . . . 67
3.9 Final remarks . . . 67

4 Reliability of LLMs in Providing Medical Advice 69
4.1 Background . . . 70
4.2 Experimental design . . . 71
4.2.1 Models . . . 71
4.2.2 Health-related questions . . . 72
4.2.3 Prompts . . . 72
4.3 Results . . . 73
4.3.1 Zero-shot experiments . . . 73
4.3.2 Few-shot experiments . . . 74
4.4 Error Analysis . . . 76
4.5 Final remarks . . . 76

5 Tools for massive processing of online content 78
5.1 Social Minder . . . 79
5.1.1 Background . . . 79
5.1.2 Architecture . . . 79
5.1.3 COVID-19 misinformation use case . . . 81
5.1.4 Final remarks . . . 82
5.2 An unsupervised perplexity-based approach for boilerplate removal . . . 83
5.2.1 Background . . . 83
5.2.2 Methodology . . . 84
5.2.3 Perplexity models . . . 85

5.2.4 Experimental settings 86
5.2.5 Results 89
5.2.6 Discussion 96
5.2.7 Python package and Web demo 96
5.2.8 Final remarks 97

6 Conclusions **99**
6.1 Future work 101

Bibliography **103**

List of Figures **121**

List of Tables **123**

Abstract

The evolution of the Web has led to an improvement in information accessibility. This change has allowed access to more varied content at greater speed, but we must also be aware of the dangers involved. The results offered may be unreliable, inadequate, or of poor quality, leading to misinformation. This can have a greater or lesser impact depending on the domain, but is particularly sensitive when it comes to health-related content.

In this thesis, we focus in the development of methods to automatically assess credibility. We also study the reliability of the new Large Language Models (LLMs) to answer health questions. Finally, we also present a set of tools that might help in the massive analysis of web textual content.

1 Introduction

The remarkable evolution of the World Wide Web and the proliferation of social networks have undeniably revolutionised the way information is accessed and disseminated, as corroborated by data extracted from a comprehensive study conducted by Reuters at the University of Oxford [145]. This transformative shift has, in numerous aspects, been advantageous, facilitating rapid access to a diverse array of content. Nonetheless, it is of utmost importance to be mindful of the potential hazards that may arise from this accessibility. The results provided by automated information retrieval tools can often be characterised as unreliable [2], inadequate [52], or low quality [146], thereby producing situations of "*misinformation*" (unintentional) or "*disinformation*" (intentional). This research thesis focuses on the first class and we mainly intend to advance in automatic classification and search methods to support end-users in their access to reliable health-related contents.

Previous empirical investigations have demonstrated that misinformation can lead to erroneous decision-making, as evidenced by the experiments of Pogacar et al. with search engine users [131]. In this seminal study, the authors showed that users who examined search results containing even partially incorrect information were considerably more susceptible to making incorrect decisions. The authors simulated different conditions: a control condition (decision without observing any search result), a search engine results page (SERP) biased towards incorrect outcomes (contradicting the medical consensus) and a SERP biased towards correct outcomes. They repeated this process for ten different medical queries, i.e. "*Do benzodiazepines help alcohol withdrawal?*", and demonstrated that with the condition biased towards incorrect results the percentage of correct decisions drops substantially.

The effect of misinformation can be more or less harmful depending on the context, but it is particularly worrying in the case of health-related information. It may even lead to personal harm if treatments are applied without proper medical supervision [162]. This became particularly evident with the COVID-19 pandemic in 2020, where misinformation about the disease and its treatments proliferated, including rumors, stigma, and conspiracy theories [84]. These researchers monitored claims about the disease and its treatments that were published in online media in the early months of 2020 and found that 82% of them contained some form of false information.

Moreover, search engines (SEs) have emerged as prevalent tools for procuring medical advice [65], and the use of social networks for accessing medical information has been increasingly growing, as indicated in the study by Reuters [145]. According to previous studies, around 72% of American internet users used the Web to find information about a health problem or treatment [66]. Another survey showed that 60% of European citizens go online to search for medical advice [45]. Regrettably, online platforms are rife with medical hoaxes, unfounded miracle diets, and recommendations published by unqualified individuals. Despite this proliferation, few solutions to automatically deal with health misinformation have been proposed [163]. As stated before, it is hazardous to adhere to recommendations posted online without proper consultation with a medical professional.

In recent years, with the proliferation of the so-called Large Language Models (LLMs), a new, more conversational form of information access has become popular [25, 62, 123]. These advanced language services arise as new reference tools to support multiple types of information needs. People can resort to LLMs to resolve medical information needs and, thus, we need to put them under severe scrutiny. A recent study [179] analysed the impact of prompts in health information seeking. However, the study was confined to a single LLM (ChatGPT) and the main goal was to evaluate prompts that incorporate supporting and contrary evidence obtained from a search engine. They reported a 80% accuracy on medical questions when using only the internal knowledge of the LLM. A drop in performance was noticed when incorporating search evidence to the prompt. The literature still lacks a comprehensive study comparing multiple models and testing their abilities to provide medical advice. There is also a need to analyse the influence of different prompts on the final performance. We will try to cover this gap in one part of our research.

In light of all these concerns, it is crucial for both users and developers of information access tools to exercise vigilance and discernment when navigating the vast landscape of online content. The onus lies with the scientific community, policymakers, and technology companies. We need to devise strategies that mitigate the dissemination of misinformation, particularly in the health domain. By fostering a culture of critical thinking, promoting digital literacy, and implementing robust fact-checking mechanisms, we can make that the benefits of increased information accessibility are not overshadowed by its potential dangers.

In line with these statements, the credibility of online information has been extensively studied. For example, Griffiths et al. [74] showed that algorithms like PageRank were unable to determine reliability of online content on their own. Ginsca and colleagues presented a thorough survey on existing credibility models from different information seeking perspectives [71]. Viviani and Pasi reviewed the main automatic methods to estimate credibility in social media, paying special attention to health content [163]. On the other hand, other researchers studied consumers' adherence to a concrete site based on their perception of credibility [116]. Easting

and colleagues also demonstrated that both the source and the knowledge of the content had a clear influence on users' perception about online health information [51].

Fogg and his colleagues defined the *prominence-interpretation theory* to determine which elements of a website influence the perception of credibility [60]. This theory was later tested through a user study involving 2,500 participants [61]. It has been proved that credibility is a subjective concept and prone to biases [88]. Researchers found evidence that, apart from the characteristics of the web elements, the receiver's characteristics also influence the perception of the information [169]. Some interesting conclusions are that subjective ratings depend on the user's background (e.g., influenced by the number of years of education or reading skills [76]).

Other lines of research follow user-oriented methods to understand how the search engine result pages (SERPs) are used to determine credibility. Kattenbeck and Elsweiler conducted a controlled study that proved that even the same snippets provided different information cues to different assessors [92]. Unkel et al. tested the importance of the ranking in the perception of credibility, demonstrating that people tend to trust the order of the results provided by the search engine [159]. Bink and colleagues went one step further and studied the influence of featured snippets on users' credibility perception. They found that users judge the information in these snippets as more credible than the rest of the results [21]. Other research initiatives focused on nudging strategies to orientate social media users towards more credible contents [20].

More in line with the main core of this research, some teams focused on assessing the credibility of health-related content on the web. For example, Matthews et al. [114] analysed a corpus about alternative cancer treatments and found that almost 90% contained false claims. Liao and Fu [99] studied the influence of age differences in credibility judgments related to health and argued that older adults care less about the content of the site. Other teams focused on how to present medical information on a search engine result page to improve credibility judgments [150].

Sondhi and his colleagues presented the first automatic approach, based on traditional learning algorithms, for medical reliability prediction at a document-level [155]. This study represents an important reference for the research presented in this dissertation. As a starting point, we tried to reproduce this research and obtained results superior to those achieved by the authors [54]. Moreover, we have extended the analysis to more collections [150, 86] and evaluated additional variants of these classic models.

In this dissertation, we present a thorough study that goes from evaluating classic approaches to the health misinformation detection problem to the study of the role of emergent LLMs in this field. We considered it critical to revisit the seminal efforts in this research field and use them as a starting point for our own research. Thus, we have reproduced Sondhi's seminal attempt to classify medical webpages in terms of reliability. Next, we proceeded to compare the capabilities of classical and neural models in identifying misinformation.

After these first steps, we propose our own multistage retrieval system for misinformation detection and compare it with state-of-the-art solution. We also dedicate some effort to put under scrutinity the reliability of the newest LLMs to provide medical advice.

Finally, we propose a set of auxiliary tools for processing massive online content. These complementary tools represent specific demonstrators or pre-processing elements that help in misinformation detection. For example, we have developed Social Minder, a platform that tracks misinformation published on online media (Twitter). We also proposed a pre-processing method for cleaning text and demonstrate its utility for several downstream tasks. These technological byproducts resulted from the scientific activities performed during these years.

It must also be noticed that, during the development of this thesis, we have participated in three editions of the Text Retrieval Conference (TREC), the prestigious evaluation campaign. More specifically, we actively participated in the Health Misinformation (HM) Track [42, 41]. This allowed us to test our methods under competitive real-world conditions. We obtained our best result, a third position, in the 2021 edition. The main goal of this track is to develop retrieval methods that promote correct and credible retrieval results over misinformation. The 2020 edition focused on misinformation related to COVID-19 and SARS-CoV-2. Our 2020 participation was exploratory and it was restricted to deploy the classifier that we replicated following Sondhi's classic approach. In 2021 and 2022, we tested our own multistage retrieval pipeline, which takes into account several signals for estimating misinformation. In 2022, a new challenge, oriented to estimate the correct answer for medical questions, was introduced. We participated into this new task with some variants consisting of exploiting a search engine's top results or prompting a LLM (GPT-3).

Information Retrieval (IR) is a multidisciplinary field that is related to various aspects of computer science, information science, and linguistics [47]. It deals with the process of retrieving relevant information from a vast collection of data sources, such as text documents, images, videos, and audio recordings. The primary objective of IR is to provide users with accurate and timely access to the information they seek.

The advent of the World Wide Web has significantly transformed the way we access and consume information. The proliferation of social media platforms and search engines has made it easier than ever before to obtain information on virtually any topic. However, this accessibility has also given rise to a significant problem: the dissemination of misinformation. In particular, health-related misinformation can be especially dangerous as it can lead people to make incorrect decisions about their health.

The problem of detecting health-related misinformation in online content has become increasingly pressing in recent years. With the rise of social media platforms and search engines as primary sources for obtaining medical advice, there has been an exponential increase in the amount of health-related content available online. Unfortunately, a large portion of this content

is inaccurate or misleading.

The detection of health-related misinformation in online content poses several challenges for researchers in the field of IR. On the one hand, there is a need for effective algorithms that can accurately identify instances of misinformation within large volumes of data. In this regard, prestigious conferences such as the TREC, organised by NIST, have carried out initiatives such as the Health Misinformation (HM) Track to encourage research on retrieval methods that promote correct and credible information versus misinformation [41, 42].

On the other hand, there is a need for robust evaluation metrics that can measure the effectiveness of these algorithms accurately. Clarke and his colleagues proposed an innovative metric: compatibility [43]. Compatibility estimates the similarity between a ranked list provided by an automatic system and an ideal ranking. Clarke and colleagues utilised Rank Biased Overlap (*RBO*) to compute compatibility between an ideal ranking (with no misinformation) and a ranking produced by a search system [170].

Several approaches have been proposed for detecting health-related misinformation in online content. These include machine learning-based methods that use natural language processing techniques to analyse text data and identify signs of misinformation [155, 135]. In this research, we present a series of contributions that go from reviewing and updating existing proposals to developing entirely new systems for health-related misinformation detection.

In conclusion, the problem of health-related misinformation detection in online content is a significant challenge for researchers in the field of IR. The development of new effective algorithms and evaluation metrics is crucial to properly address this problem. By mitigating the dissemination of health-related misinformation, we help individuals to have access to accurate and reliable information.

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that deals with the interaction between computers and humans in natural language. Over the years, NLP has evolved significantly, with the development of a wide range of models that aim at enabling computers to understand and generate human language. The origins of this discipline go back to the early 1950s with Chomsky's seminal work on methodologically translating natural language sentences into a computer-understandable format [36]. Until the 1980s, most NLP systems were based on complex systems of rules [69]. In contrast, from the 1990s onwards, statistical modelling became mainstream in this area. The main goal of statistical language modelling is to learn probability functions of sequences of words in a given language. For example, distinct approaches have been proposed to represent *n-grams* [35].

In 2001, Bengio and his colleagues initiated a paradigm shift by proposing an efficient neural language model (LM), which exploits a feed forward network [16]. The goal was to fight the curse of dimensionality in traditional language modelling (i.e., a test sequence is likely to be different from all sequences seen at training time). Bengio's core idea consisted of learning

distributed representations of words. In this way, each training sentence can inform the model about a large number of semantically similar sentences.

Another milestone in the development of LMs occurred with the appearance of Word2Vec [117]. This model was introduced in 2013 as a neural network-based model that aimed to learn continuous vector representations of words from large text corpora. The model considers two architectures: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts a word based on its context words, while Skip-gram predicts context words based on a given word. Word2Vec was able to capture semantic and syntactic relationships between words and achieved state-of-the-art performance on various NLP tasks.

In 2014, GloVe was designed to address some of the limitations of Word2Vec [128]. While Word2Vec focused on local context, GloVe aimed to capture global word co-occurrence statistics. The model learns word embeddings by minimising the difference between the dot product of embeddings and the logarithm of the word co-occurrence probability. GloVe achieved better performance than Word2Vec on some NLP tasks.

In recent years, Long Short-Term Memory (LSTM) networks gained popularity for NLP tasks. LSTMs are a type of Recurrent Neural Network (RNN) that can capture long-term dependencies in sequential data by selectively forgetting or remembering information. However, LSTMs have some limitations, such as difficulty in parallelisation and the inability to capture global dependencies effectively [70].

To overcome these limitations, attention mechanisms were introduced. This allows models to focus on specific parts of the input sequence while processing it [161]. Attention mechanisms have been used successfully in various NLP tasks such as machine translation and text summarisation. These models, called Transformers, quickly became popular due to their ability to parallelise computations efficiently and capture global dependencies effectively. Some examples of this architecture are BERT [50] and the sequence-to-sequence model, T5 [143].

Transformer-based architectures represent advanced devices that are promising for supporting the detection of misinformation within textual extracts. In this thesis, we will evaluate their capabilities against more traditional models and use them as essential parts of a multistage retrieval system. The idea is to exploit these sophisticated language-based artifacts to support tasks such as the identification of correct and credible information and the filtering of health-related misinformation.

A highly significant breakthrough in recent years has been the development of Large Language Models (LLMs), such as GPT-3 or GPT-4 [25, 123]. These models are also based on the transformer architecture. However, they differ from others, such as BERT, in that they only implement the decoder stack. Thus, we refer to them as *decoder-only models*. The choice of prioritising the encoder or decoder blocks in the architecture is directly related to the selection of pre-training objectives. GPT-based models focus on autoregression objectives or regenerating

missing text sequences. They are particularly skilful in text generation tasks such as machine translation, summarisation and question-answering.

In November 2022, the release of OpenAI's Chat-GPT[1] marked a significant milestone in the evolution of LLMs. This model is fine-tuned with human feedback using reinforcement learning [125]. Its appearance represented a new paradigm shift towards conversational information access powered with excellent AI capabilities. However, there are still some open research questions regarding these models. In concrete, its knowledge and their ability to provide correct medical advice must be put under scrutiny and, thus, we have designed some research activities to address this goal.

1.1 OBJECTIVES

The general goal of this PhD dissertation is to advance in automatic classification and search methods to support end-users in their access to reliable health-related contents. In particular, we want to address the following scientific challenges:

- O1. Evaluate the usefulness of **classical machine learning approaches** to the health misinformation detection problem. More specifically, test the ability of classic text classification methods to support document-level predictions of reliability. This will include activities such as replicability studies of results published by reputed research teams, comparisons of content-based, link-based and other document features as predictor variables, and transfer learning experiments to understand the applicability of this classic technology to other external collections (including data from well-known evaluation campaigns).

- O2. Evaluate the **new transformer models** for document-level reliability estimation and compare them with more classical solutions. To that end, we will consider different relevant target variables, such as trustworthiness and readability, and compare the variants under different conditions of training data.

- O3. Test the combination and predictive power of **multiple sources of evidence** for the health misinformation detection problem. Understand which sources of evidence or signals can help determine the reliability of online health-related information and how to combine these pieces of evidence. We will consider a wide range of features, including estimates at document and passage level, supervised and non-supervised signals powered by Deep Learning technology, re-ranking stages and different forms of fusion. The signals range from traditional relevance estimators, such as BM25, to sophisticated reliability predictors based on emergent Transformer architectures.

[1] https://chat.openai.com/chat

– O4. Evaluate the new **Large Language Models** and test their abilities in providing **reliable medical advice**. With the current shift from classic search engines to advanced conversational AI agents, it becomes critical and socially relevant to determine the capabilities of the emergent LLMs in providing correct medical advice. We want to test different contextual settings and determine their influence in the quality of the results. We also intend to qualitatively evaluate these models and understand their errors and potential downsides.

– O5. Develop a set of complementary **tools** that aid in the **processing of massive social data**. As part of this research, we want to design and develop new architectures and tools capable of processing massive amounts of text published on social media. This will allow to test and deploy the research methods developed in this thesis under relevant use cases (for example, leading to demonstrators of misinformation detection in certain domains). This also involves the development of appropriate NLP pre-processing tools that give support in multiple downstream tasks, including misinformation detection.

1.2 Methodology

In this thesis we use the typical guidelines of research projects based on the scientific method. We will follow an iterative process to advance in automatic classification and search methods to support end-users in their access to reliable health-related contents. First of all, we take as starting point some reputed studies on the field of reliability prediction for health-related content. Replicability of previous experiments is crucial in the advancement of science and, thus, we seriously take replicability of existing methods as core elements in our research. Next, we critically evaluate the capabilities of traditional learning methods and compare them against newer neural-based approaches for misinformation detection. As a matter of fact, a fundamental goal of information access research is to develop new techniques, and to demonstrate that they attain improved effectiveness compared to their predecessors.

We also propose our own multistage retrieval system for misinformation detection. In this step, we evaluate the inclusion of different sources of evidence to identify misinformation and we assess their importance in the final classification result. To that end, we design a comprehensive set of empirical tests, including ablation experiments.

For the evaluation of results, the field of IR has clearly defined methodologies. The existence of standardised test collections, evaluation metrics, and so forth stand on the foundations of the IR field. From a methodological viewpoint, we will carefully follow the existing evaluation standards, resort to validated benchmarks and compare the results against competitive alternatives.

We face an innovative target application and, thus, it is pivotal to be able to quantify the quality of health-related information (e.g. COVID-19 treatments) accessed online. This introduces the need of expanding the toolkit of effectiveness metrics. Related to this, we follow here recent advances in the field such as compatibility measures. This will be complemented with more standard classification and search metrics, such as F1, Precision, Recall or NDCG. Furthermore, we also test the quality of our solutions through the participation in the prestigious TREC conference, more specifically in its Health Misinformation Track. The active participation in well-known evaluation campaigns is a valuable asset for this thesis.

On the other hand, this research also focuses on emerging ways to access information. We study in-depth the capabilities of the newest LLMs to provide correct medical advice under different contexts. To that end, we follow recent trends on prompt engineering and in-context learning and compare multiple LLMs on a wide range of health information needs.

In parallel to the scientific activities, this research had a strong technological component. New technologies were developed to enable mass analysis of natural language in web media (e.g., social networks or conventional websites). Accessing and processing online information is a challenge in itself, and being able to do so in real-time is an added value for many applications. For example, it is necessary to remove all documents that are irrelevant to our task (e.g., content posted online that is not related to health). In addition, flexible and efficient solutions have to be developed for the massive processing of web data in real time. This involves the construction of a new Big Data platform that natively runs parallel applications in languages such as Python or Perl. Related to this, some pre-processing tools that can boost the performance of downstream tasks were proposed. To sum up, an incremental working approach, based on agile methodology concepts, was used. A Kanban board was used to prioritise tasks, which were evaluated in weekly meetings with the thesis supervisors.

1.3 PUBLICATIONS

The research developed in this thesis has led to the following contributions:

Journals:

Fernández-Pichel, M.[a], Losada, D.E.[a], and Pichel, J. C.[a]. (2022). *A multistage retrieval system for health-related misinformation detection*. Engineering Applications of Artificial Intelligence, 115, 105211. The publication is available at: `https://doi.org/10.1016/j.engappai.2022.105211`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: Journal Impact Factor: 8.0 (JCR 2022), 1.73 (SJR 2022). Journal ranked in JCR 2022, in Computer Science, Artificial Intelligence, (Q1, 25/145).

– **PhD candidate contribution**: Research conceptualization, design and implementation of the system, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: Elsevier allows its inclusion as a part of doctoral thesis without express permission (see `www.elsevier.com/about/policies/copyright#Author-rights` or Figure 1.1)

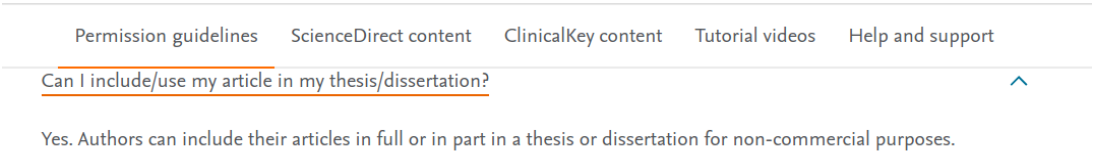


Figure 1.1: Licensing information of the previous Elsevier publication.

> Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], Pichel, J. C.[a], and Gamallo, P.[a]. (2023). *An unsupervised perplexity-based method for boilerplate removal*. Natural Language Engineering, 1–18. The publication is available at: `https://doi.org/10.1017/S1351324923000049`
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: Journal Impact Factor: 2.5 (JCR 2022), 0.58 (SJR 2022). Journal ranked in JCR 2022, in Linguistics, (Q1, 33/194).

– **PhD candidate contribution**: Research conceptualization, design and implementation of the tool, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: This article has been published Open Access (see Figure 1.2).

Figure 1.2: The previous paper has been published as open access.

Conferences:

Fernández-Pichel, M.[a], Losada, D.E.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. (2021). *Reliability Prediction for Health-Related Content: A Replicability Study*. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham. The publication is available at: `https://doi.org/10.1007/978-3-030-72240-1_4`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

- **Quality indicators**: Conference ranked as Class 2 in GII-GRIN-SCIE (GGS) Conference Rating and also ranked as CORE:A.
- **PhD candidate contribution**: Research conceptualization, experimentation, and partially manuscript writing.
- **Reproduction rights**: *Reproduced with permission from Springer Nature*, see Figure 1.3.

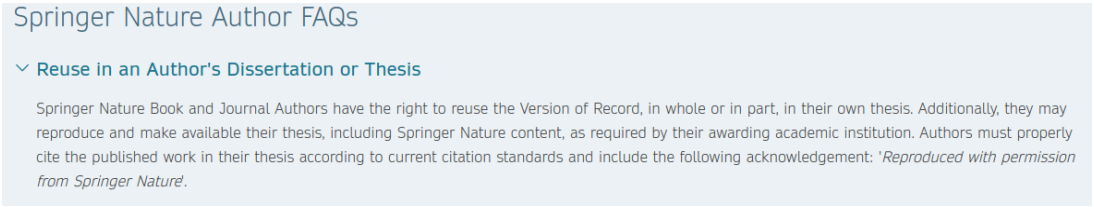
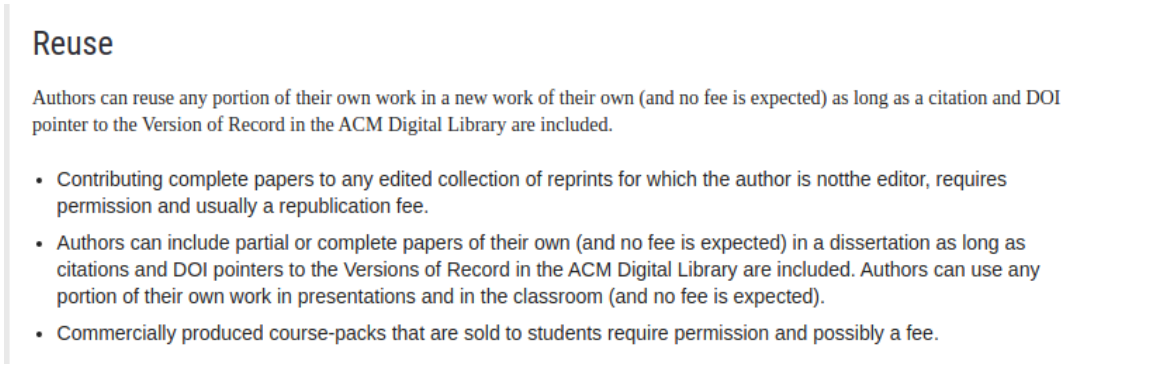


Figure 1.3: Licensing information of a Springer publication.

Fernández-Pichel, M.[a]. (2021). *Estimating the Reliability of Health-related Search Results*. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2702. The publication is available at: `https://doi.org/10.1145/3404835.3463266`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

- **Quality indicators**: Conference ranked as Class 1 in GII-GRIN-SCIE (GGS) Conference Rating and also ranked as CORE:A++.
- **PhD candidate contribution**: Research conceptualization, experimentation, and partially manuscript writing.
- **Reproduction rights**: see Figure 1.4.



Reuse

Authors can reuse any portion of their own work in a new work of their own (and no fee is expected) as long as a citation and DOI pointer to the Version of Record in the ACM Digital Library are included.

- Contributing complete papers to any edited collection of reprints for which the author is notthe editor, requires permission and usually a republication fee.
- Authors can include partial or complete papers of their own (and no fee is expected) in a dissertation as long as citations and DOI pointers to the Versions of Record in the ACM Digital Library are included. Authors can use any portion of their own work in presentations and in the classroom (and no fee is expected).
- Commercially produced course-packs that are sold to students require permission and possibly a fee.

Figure 1.4: Licensing information of an ACM publication.

Fernández-Pichel, M.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. (2020). *CiTIUS at the TREC 2020 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec29/papers/CiTIUS.HM.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

– **Quality indicators**: Conference ranked as Class 2 in GII-GRIN-SCIE (GGS) Conference Rating (This was the class in the GGS Conference Rating assigned to this conference at the time of publication, period 2018-2021).

– **PhD candidate contribution**: Research conceptualization, design and implementation of the algorithms, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: Free available for the research community by NIST.

Fernández-Pichel, M.[a], Losada, D.E.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. (2021). *Comparing Traditional and Neural Approaches for Detecting Health-Related Misinformation*. In: Candan,S. , et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, vol 12880. Springer, Cham. The publication is available at: `https://doi.org/10.1007/978-3-030-85251-1_7`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

– **Quality indicators**: Conference ranked as Class 3 in GII-GRIN-SCIE (GGS) Conference Rating.

– **PhD candidate contribution**: Research conceptualization, experimentation, and partially manuscript writing.

– **Reproduction rights**: *Reproduced with permission from Springer Nature*, see Figure 1.3.

Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], Pichel, J. C.[a], and Gamallo, P.[a] (2021). *CiTIUS at the TREC 2021 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec30/papers/CiTIUS-HM.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: Conference ranked as Class 3 in GII-GRIN-SCIE (GGS) Conference Rating (This conference was downgraded in GGS Conference Rating 2021 to Class 3).

– **PhD candidate contribution**: Research conceptualization, design and implementation of the system, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: Free available for the research community by NIST.

Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], and Pichel, J. C.[a] (2022). *CiTIUS at the TREC 2022 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec31/papers/CiTIUS.H.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: Conference ranked as Class 3 in GII-GRIN-SCIE (GGS) Conference Rating (This conference was downgraded in GGS Conference Rating 2021 to Class 3).

– **PhD candidate contribution**: Research conceptualization, design and implementation of the system, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: Free available for the research community by NIST.

Fernández-Pichel, M.[a], Losada, D. E.[a], and Pichel, J. C[a]. (2022). *Social Minder: a Tool for Social Media Monitoring and its Use for Detecting COVID-19 Misinformation*. In 2nd Joint Conference of the Information Retrieval Communities in Europe, CIRCLE 2022. CEUR Workshop Proceedings. The publication is available at: `https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_01.pdf`.

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: -

– **PhD candidate contribution**: Research conceptualization, design and implementation of the system, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: The reproduction rights of this publication have been granted by the CC BY licence, see Figure 1.5.

Figure 1.5: Licensing information of a CEUR-WS publication (CIRCLE 2022).

> Fernández-Pichel, M.[a], Losada, D. E.[a], and Pichel, J. C[a]. (2020). *eXtream: a System for Real-time Monitoring of Dynamic Web Sources*. In 1st Joint Conference of the Information Retrieval Communities in Europe, CIRCLE 2021. CEUR Workshop Proceedings. The publication is available at: `https://ceur-ws.org/Vol-2621/CIRCLE20_34.pdf`.
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: -

– **PhD candidate contribution**: Research conceptualization, design and implementation of the system, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: The reproduction rights of this publication have been granted by the CC BY licence, see Figure 1.6.

Figure 1.6: Licensing information of a CEUR-WS publication (CIRCLE 2020).

Fernández-Pichel, M.[a], Meyer, S.[b], Bink, M.[b], Frummet, A.[b], Losada, D. E.[a], and Elsweiler, D. [b]. (2023). *Improving the Reliability of Health Information Credibility Assessments*. In Proceedings of the 3rd Workshop on Reducing Online Misinformation through Credible Information Retrieval 2023 (ROMCIR 2023), Dublin, Ireland. The publication is available at: `https://ceur-ws.org/Vol-3406/paper4_jot.pdf`.

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

– **Quality indicators**: -

– **PhD candidate contribution**: Research conceptualization, design and implementation of the guidelines, evaluation and verification, and partially manuscript writing.

– **Reproduction rights**: The reproduction rights of this publication have been granted by the CC BY licence, see Figure 1.7.

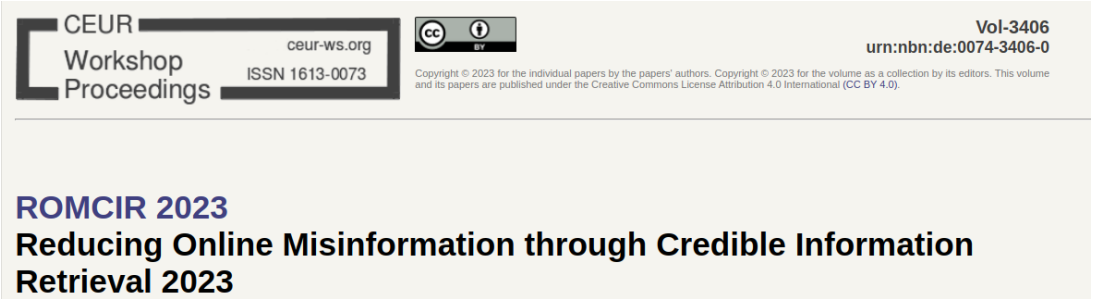


Figure 1.7: Licensing information of a CEUR-WS publication (ROMCIR 2023).

Other journal publications not related to this thesis:

Fernández-Pichel, M.[a], Ezra-Aragon, M.[a], Saborido-Patiño, J.[a], Losada, D.E.[a] (2023). Personality Trait Analysis during the COVID-19 Pandemic: a Comparative Study on Social Media. Journal of Intelligent Information Systems. 1–6, Springer US. The publication is available at: `https://doi.org/10.1007/s10844-023-00810-3`.

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

– **Quality indicators**: Journal Impact Factor: 3.4 (JCR 2022), 0.8 (SJR 2022). Journal ranked in JCR 2022, in Computer Science, Artificial Intelligence, (Q3, 77/145).

– **PhD candidate contribution**: Research and conceptualization, data curation, validation, writing-review and editing.

– **Reproduction rights**: *Reproduced with permission from Springer Nature*, see Figure 1.3.

1.4 Dissertation structure

The structure of this PhD dissertation is organised in seven chapters. This document covers in detail the state-of-the-art, the research challenges and objectives of the thesis, the design and development of the proposed solutions, a discussion and analysis of the experimental results, and finally the conclusions derived from this work. Specifically, we structured this document as follows:

- An introduction about the health misinformation detection problem together with a brief review of the state-of-the-art was presented in Chapter 1. In particular, we briefly introduced the IR and NLP fields, their close relationship and their connection with our research problem. The objectives of this PhD dissertation, methodology and list of publications were also detailed in this chapter.

- Chapter 2 focuses on presenting some seminal approaches to the problem. We conduct a replicability study, working with a previous work on reliability classification of medical web pages and obtaining better performance than that achieved by the original study. Next, we carried out a comparison between traditional and neural approaches for detecting health misinformation online. Finally, we also report our first participation in the TREC 2020 HM track.

- In Chapter 3, we present the core system of this research: a multistage retrieval system for misinformation detection. We detail the multistage approach and all the implementation and experimentation details, showing that it obtains competitive results. In this same chapter, we report our participation in the TREC 2021 and TREC 2022 HM track, using the multistage system previously described.

- In Chapter 4, we go one step beyond and, instead of any custom solution, we evaluate the capabilities of the new LLMs to provide medical advice.

- Chapter 5 describes some research lines that emerged in parallel to the development of this thesis. First, we present a Big Data system for the massive processing of data published on social media and their subsequent evaluation of misinformation detection. Second, we describe a new text preprocessing technology that estimates the presence of malformed text, a critical task to obtain good performance in any downstream task.

- Finally, Chapters 6 presents some final conclusions and lines of future work.

2 Classic Approaches to the Health Misinformation Detection Problem

Although classical approaches are expected to be surpassed by more recent methods, it is important to consider them as a reference in the study of methods for detecting health-related misinformation. A seminal attempt to automatically classify reliable medical webpages was done by Sondhi's et al [155], and we have considered this study as a core reference in the empirical evaluation of classic methods.

In Section 2.1, we detail our efforts to reproduce Sondhi's findings in an attempt to compare traditional feature variables and classification methods. The ability to replicate previously published research is crucial for the advancement of science. Furthermore, we have substantially extended the original empirical study, by including additional datasets, and report performance results that are superior to those achieved by the original work. We then deployed these classification methods in the context of our participation in the TREC 2020 Health Misinformation Track. We detail our experience in this well-known shared-data experimental campaign. Our participation in this renowned evaluation initiative helped to further test the the generalisation capabilities of these classical methods.

As a natural next step, we proceeded to compare the abilities of classical and neural models for automatically detecting misinformation. These endeavours are reported in Section 2.2. To that end, we have assessed multiple dimensions, including trustworthiness and readability, and compared models with varying percentages of training data and training time.

The contents of this chapter were extracted from the following publications:

Fernández-Pichel, M.[a], Losada, D.E.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. *Reliability Prediction for Health-Related Content: A Replicability Study*. In: Hiemstra, D., Moens, MF., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657. Springer, Cham. The publication is available at: `https://doi.org/10.1007/978-3-030-72240-1_4`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

Fernández-Pichel, M.[a], Losada, D.E.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. (2020). *CiTIUS at the TREC 2020 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec29/papers/CiTIUS.HM.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

Fernández-Pichel, M.[a], Losada, D.E.[a], Pichel, J. C.[a] and Elsweiler, D.[b]. (2021). *Comparing Traditional and Neural Approaches for Detecting Health-Related Misinformation*. In: Candan,S. , et al. Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2021. Lecture Notes in Computer Science, vol 12880. Springer, Cham. The publication is available at: `https://doi.org/10.1007/978-3-030-85251-1_7`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.
[b]Chair for Information Science, Regensburg University, Regensburg, Bavaria, Germany.

2.1 Reliability Prediction for Health-related Content: A Replicability Study

In this section, we report on our efforts to replicate the predictive technology developed by Sondhi and colleagues [155], based on Natural Language Processing (NLP) and Machine Learning techniques. We chose this study because, to the best of our knowledge, it represents the first attempt to address the issue of automatically assessing the reliability of webpages in the medical domain. These researchers considered this problem as a binary classification task and constructed a test dataset that is publicly available. The main goal of the paper was to compare a set of webpage features and understand how they work as predictors for the classification of the webpages as reliable or non-reliable (see subsection 2.1.1).

If we succeed in recreating their results, the conclusions extracted in the original study would be verified and reinforced. This replication effort is worthwhile to establish the utility of this classic technology and to analyze its potential to be applied in filtering non-reliable content.

To this end, we examined and, where possible, re-implemented the features proposed by the original study. In order for the results to be comparable, we applied the same experimental methodology and performance metrics proposed in the original paper. We also include a subsection in which the experiments are extended and applied to two new datasets [86, 150]. This helps to analyze the generability of the variants tested.

In the literature, researchers have employed several notions that are intimately related such as *reliability* [155], *trustworthiness* [86], *credibility* [150], or *veracity* [165]. Following the study that we take as our main reference for replicating classical techniques [155], we adopt the notion

	Webpages
# Reliable	180
% Reliable	50%
# Unreliable	180
% Unreliable	50%

Table 2.1: Class distribution in Sondhi's dataset.

of reliability. For determining reliability, Sondhi and colleagues defined a set of guidelines using the eight HONcode Principles[1]. The first part of our replicability study focuses on predicting the reliability of the webpages from the original corpus, while the additional experiments with other datasets consider other notions, such as credibility or trustworthiness, as proxies of reliability (see subsection 2.1.4).

2.1.1 Dataset

Sondhi et al. manually created a **fully balanced** dataset with reliable and unreliable webpages (see Table 2.1) that we directly used in our replicability study. This eases the classification task, but it is not very realistic since real-world applications rarely face situations where the ratio of cases is the same for both classes.

In the original paper, the authors randomly selected the positive pages from those websites accredited by HON[2] according to their principles. On the other hand, as HON does not report non-accredited sites, they searched the Web with a deliberate strategy to find poor quality pages. To that end, they employed hand-crafted queries, such as *disease name + "miracle cure"*. To ensure a topical overlap between negative and positive instances (i.e. to avoid topic-bias classification), they conducted a topic analysis over the reliable corpus and extracted keywords related to diseases that occur in the set of reliable pages. For each keyword, they manually produced queries which involved terms like *treatment* or *miracle*. Finally, the authors checked and selected 180 unreliable pages from the search results. As the original download link for the dataset was no longer valid, the dataset was sourced via personal communication with the authors.

The main goal of the original paper was to build a **document-level classifier** using a standard supervised learning approach. We followed their experimental setup, in which the authors argued that reliability can be represented as a binary variable.

2.1.1.1 Features

A variety of **features** were proposed based on style, content and external information such as links. As will be seen, we were not able to apply all of these in our experiments, since some

[1] https://www.hon.ch/cgi-bin/HONcode/principles.pl?English
[2] https://www.hon.ch/en/

tools or libraries were outdated, and other elements were not described in a sufficiently detailed manner. In the original paper, webpages were represented using several features, namely:

– **Link-based features**: the number and type of links are usually a good indicator of the type of website we are dealing with [14, 23]. For example, as Sondhi and colleagues argued, a more reliable site tends to have more internal links, while a less reliable site tends to have more external links and advertisements [177]. On the other hand, the presence or absence of privacy policy information or contact links for the page author can be indicators of reliability. This is because the presence of these types of elements gives a sense of confidence to the user who consults the resource [72, 89].

  Based on these criteria five features were defined: normalised value of internal links, normalised value of external links, normalised value of total links, the presence or not of contact link (boolean), and the presence or not of privacy link (boolean). For the latter two, the original paper did not explain how they were computed. Therefore, we manually defined two lists of privacy[3] and contact[4] expressions, such as *Privacy Policy* or *Contact Us*, after performing a first exploratory analysis over the documents.

  For normalisation purposes, the original authors analysed a random sample of documents and they experimentally chose a large normalisation denominator (the link count was divided by Z_1, which was set to 200).

  In our experiments, the links were extracted from the text using the Beautiful Soup[5] Python package.

– **Commercial features**: the presence of commercial interest and advertisements often indicates a low reputation [14, 177]. Therefore, two commercial-based features were defined: the normalised value of commercial links and the normalised frequency of commercial words in the website.

  For the latter, an initial list of indicative words of commercial interest was proposed in the article. We manually extended this list[6]. Since the original article was not explicit about word preprocessing, we followed a naive approach in which a word must match exactly with some of the words in the list to be taken into account in the final metric. This strategy could be improved in future experiments by applying lemmatization techniques, for example.

[3] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/privacy.txt
[4] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/contact.txt
[5] https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[6] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/comm_list.txt

Regarding normalisation, the normalised value of commercial links was obtained dividing by the same Z_1 used above. The second feature consisted of dividing the number of commercial words found by the document's length.

- **PageRank Features**: the authors of the original paper used this feature as an indicator of the relative importance of a website [5]. However, this service has been removed by Google, and all Python packages that used their endpoint are not functioning. It would be still possible to manually compute PageRank based on the web graph. However, the current web graph does not reflect the situation of these pages when the collection was created (some pages are no longer accessible). Furthermore, previous work has shown that such features capture the popularity of a website, but fail to measure reliability [134].

- **Presentation features**: reliable content is usually presented carefully and clearly [72]. To evaluate this, the original paper employed *elinks*[7], a tool to extract the text of the webpage. Then, they defined two features based on the number of blank lines. However, in the final comparison, Sondhi et al. did not include this feature set and, thus, we have not incorporated these features into our replicability effort.

- **Word-based features**: textual content and style are often good indicators of the reliability or reputation of a website [113, 120]. Therefore, each word in a document was considered as a different dimension, represented by its normalised frequency. Since the authors did not declare the use of any preprocessing stage, we applied no stemming or lemmatization.

  We additionally considered two alternative pre-processing strategies, with and without *stopword* removal. To support this stage, the NLTK[8] English *stoplist* was manually extended with additional common words [9]. This was done after a preliminary exploration of the documents.

  Finally, for each word we divided the number of occurrences of the word by the document length.

In addition to testing the feature sets in isolation, Sondhi and his team also considered a final combination that merged **all features together**. In our case, we tested two variants of "all features" (one with word features extracted with stopword removal and another one with word features extracted with no stopword removal).

2.1.2 Experimental setup

[7] http://elinks.or.cz
[8] https://www.nltk.org/nltk_data
[9] https://github.com/MarcosFP97/Health-Rel/blob/master/lexicon/stopwords.txt

	Weighted Accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	60.8	71.1	79.6
Links + Commercial	67.8	75.9	79.6
Words	**80.6**	**83.9**	85.0
All	80.0	83.2	**86.8**

Table 2.2: Sondhi et al. original paper results.

When carrying out the experimentation, a **vector support machine**, which supports two-class classification, was used as learning method. The original paper used a C++ implementation but, for compatibility reasons, we employed the SVMlight[10] Python wrapper.

To evaluate the results, we applied **5-fold cross validation**, as in the original study. When generating the predictions, there could be **two types of errors**: classifying a reliable page as non-reliable (FP) and classifying a non-reliable page as reliable (FN). The latter being the one we wish to avoid most. To make results comparable, the performance metric used is the same as in the original paper:

$$\text{Weighted Accuracy}(\lambda) = \frac{(\lambda \times TP) + TN}{\lambda \times (TP + FN) + TN + FP} \tag{2.1}$$

Three variants were considered, corresponding to $\lambda \in \{1, 2, 3\}$. Moreover, following the original paper, the SVM classifier was trained with a cost-factor set to the value of λ (the weighted accuracy λ=1 was obtained with a SVM whose cost-factor was set to 1, the weighted accuracy λ=2 was obtained with a SVM whose cost-factor was set to 2, and so forth). Such an approach tunes the classifier to the measure that would later evaluate its effectiveness.

We note that the experiments were performed on an Ubuntu 19.04 machine, with 32GB of RAM, 240GB of storage and an Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz. The Python version used was 3.7.3 in an Anaconda 4.8.0 environment. However, for the CLEF eHealth dataset experiments, detailed in subsection 2.1.4.4, it was necessary to use a server due to the storage requirements. More specifically, we used a CentOS 7.6.1810 machine, with 377GB of RAM, 15T of storage and Intel(R) Xeon(R) CPU E5-2630 v4 processor. The Python and Anaconda versions used were the same as in the local experiments.

2.1.3 Results

Sondhi et al.'s original results are shown in Table 2.2. In our experiments, we considered two variants for word-based representation: with and without *stopword* removal. Moreover, commercial features were not tested in isolation, but combined with link-based features. This is reasonable since they are intimately related to external and advertising links.

[10] https://bitbucket.org/wcauchois/pysvmlight

	Weighted Accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	70.5	80.0	73.5
Links + Commercial	69.7	79.4	74.3
Words (removing stopwords)	80.8	80.2	80.3
Words (keeping stopwords)	82.8	85.6	88.5
All (removing stopwords)	**97.5**	**98.3**	**98.6**
All (keeping stopwords)	96.1	96.3	96.5

Table 2.3: Our results for Sondhi et al. dataset.

Our results (see Table 2.3) differ from the original ones, but the same conclusions can be drawn: word-based features and the merging of all features achieve the best performance. Our comparison of the two word-based variants (with and without *stopwords*) suggests that keeping *stopwords* is the safest approach to estimate the reliability of a webpage.

We note that our best performance is higher than that obtained in the original work. More specifically, in our case, we observed a high increase in the performance obtained by merging all features together. This contrasts with the original study, where the combination of features did not add value. This is perhaps the most surprising outcome of the replicability experiments, and the only plausible explanation we can derive is that this might result from differences in setup between our experiments and the original evaluation, as described in the previous sections.

2.1.4 Additional Experiments

To build on Sondhi et al.'s work and to determine how general their findings are, we applied new **standardisation** techniques to the Sondhi et al. dataset and also tested the methods with two **additional datasets**.

2.1.4.1 Standardisation

In the original paper the authors did not report details about **standardisation** of the features. These methods are commonly applied in machine learning [79] and could affect performance. Therefore, we tested the effect of standard scaling (to get 0 mean and 1 standard deviation) and report here the results (see Table 2.4).

As can be seen, the performance of all feature sets increases in comparison with the results reported in Table 2.3. Of particular note, the models with word-based representation are most improved. By carrying out this procedure, in addition to the Z_1 normalisation per document previously described, we are favouring features or words that have a low average, that is, less-common or technical words (see Figure 2.1). This evens out the differences between terms and the classifier gives special emphasis to features that deviate from its average in a particular document (for example, a word that is broadly used). This also explains why the best feature combination is word-based with *stopwords*.

	Weighted Accuracy (%)		
Features	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	74.4	78.1	76.4
Links + Commercial	73.3	76.5	79.9
Words (removing stopwords)	97.2	**98.3**	98.5
Words (keeping stopwords)	**98.1**	**98.3**	**98.9**
All (removing stopwords)	97.2	**98.3**	98.5
All (keeping stopwords)	97.8	**98.3**	**98.9**

Table 2.4: Our results for Sondhi et al. dataset (with standard scaler).

$$\begin{array}{c} \\ D1 \\ D2 \\ \vdots \\ Dm \end{array} \begin{array}{c} \begin{array}{ccc} the & \cdots & hydroxychloroquine \end{array} \\ \begin{pmatrix} 0,6 & \cdots & 0,1 \\ 0,7 & \cdots & 0,2 \\ \vdots & \ddots & \vdots \\ 0,8 & \cdots & 0,3 \end{pmatrix} \end{array} \xrightarrow{\frac{x-\mu}{\sigma}} \begin{array}{c} \\ D1 \\ D2 \\ \vdots \\ Dm \end{array} \begin{array}{c} \begin{array}{ccc} the & \cdots & hydroxychloroquine \end{array} \\ \begin{pmatrix} 0,07 & \cdots & 0,1 \\ 0,07 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0,13 & \cdots & 0,1 \end{pmatrix} \end{array}$$

Figure 2.1: Document-term matrix standardisation.

2.1.4.2 New test datasets

The Web Search dataset by Schwarz et al. [150] and the CLEF eHealth consumer health search task (2018) [86] were used to further evaluate this classification technology. Both contain health-related content, but the first additionally addresses topics such as finance, politics, environment, and news about famous people.

Schwarz et al.'s work focused on credibility assessments and how to help people searching for information online. The CLEF eHealth task addresses a similar problem, but it is tighter to health-related online data. It must be noticed that these documents were not labelled in terms of reliability, but the notions of credibility and trustworthiness were used instead. However, we considered these concepts as proxies of reliability and studied how generalisable the previous conclusions were against these test sets.

Schwarz et al. chose 1000 webpages related to multiple topics to be labelled in terms of credibility. They proposed a five-point Likert scale, from 1 to 5, to generate the ground-truth, and one of the authors of the paper rated the whole collection.

On the other hand, the CLEF eHealth consumer health search task dataset was created from CommonCrawl webpages[11]. The organisers of the task defined an initial list of potentially interesting sites and then, they submitted queries against a search engine to retrieve the final URLs. The initial list was extended by manually adding some reliable sites and other known to be unreliable. Finally, the corpus was divided into folders by domain.

In this CLEF task, it was decided to implement the RBP-based method proposed by Moffat et al. [118] to generate the assessment pool, instead of using a fixed-depth pooling strategy. After the pool was formed, human assessors from Amazon Mechanical Turk, with certain profiles, were selected. In the case of trustworthiness judgements, an eleven point scale, from 0 to 10, was used.

[11] http://commoncrawl.org

	Schwarz et al.	CLEF eHealth
# Reliable	75	9,879
% Reliable	93.75%	73.25%
# Unreliable	5	3,607
% Unreliable	6.25%	26.75%

Table 2.5: Class distribution in the different datasets.

It was necessary to relabel both datasets into a binary scale to fit with our 2-class technology. To that end, we removed the middle values (3 for Schwarz et al. and from 4 to 6 for CLEF) and mapped the extreme values to reliable and unreliable, respectively.

The main statistics of these datasets after performing this relabelling process are shown in Table 2.5. In both cases, we face an **imbalanced data** problem. This is particularly acute in the case of the Schwarz et al. data.

Imbalanced learning is a common problem and there are multiple techniques to deal with it. In this case, we considered and compared two different approaches: introducing a **cost-factor** that applies a higher penalty to errors in the minority class and **resampling techniques** that try to balance the data by adding artificial instances or by removing some majority examples [77, 81, 80, 33]. In this report, only cost-factor techniques are shown since our preliminary experiments suggested that cost-factor methods outperform resampling methods in both datasets.

On the other hand, in imbalanced learning, it is common to use metrics, such as the **F1 measure**. Here, we report the micro-averaged F1, weighted by the frequency of each class, and the value of F1 for each class. At the time of selecting the best feature combination for each collection, we gave priority to the minority class (non-reliable F1).

Finally, it is worth noting that for both datasets the standardisation method described in 2.1.4.1 was applied.

2.1.4.3 Schwarz et al. results

Due to the small dataset size, a stratified **2-fold cross validation** was used (instead of 5-folds). The obtained results are shown in Table 2.6. We note that in case of a tie, we always selected the simplest feature set.

With **cost factor set to 1**, link-based features perform the best, but the classifier does not detect a single unreliable document. With this learning strategy, no combination is capable of correctly cataloguing examples from the minority class. This is not surprising given the low percentage of negative examples (6.25%).

With **cost factor 2**, the results were still even, but some feature combinations were able to detect the minority class. This was the case of the word-based model and for the model combining all features- keeping *stopwords*. The latter was selected as the best combination, due to a slight difference in the weighted accuracy performance.

Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	Weighted Accuracy (%) $\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	1	**0.94**	**0.97**	**0**	**93.75**	-	-
	2	0.94	0.97	0	-	88.26	-
	3	0.94	0.97	0	-	-	83.4
Links + Commercial	1	**0.94**	**0.97**	**0**	**93.75**	-	-
	2	0.94	0.97	0	-	88.26	-
	3	0.94	0.97	0	-	-	83.4
Words (removing stopwords)	1	0.93	0.96	0	92.5	-	-
	2	0.91	0.95	0.25	-	87.01	-
	3	**0.91**	**0.95**	**0.33**	-	-	**85.42**
Words (keeping stopwords)	1	0.91	0.95	0	91.25	-	-
	2	0.91	0.95	0	-	85.88	-
	3	0.91	0.95	0.2	-	-	84.54
All (removing stopwords)	1	**0.94**	**0.97**	**0**	**93.75**	-	-
	2	0.91	0.95	0	-	85.88	-
	3	0.91	0.95	0	-	-	81.13
All (keeping stopwords)	1	0.93	0.96	0	92.5	-	-
	2	**0.91**	**0.95**	**0.25**	-	**87.02**	-
	3	**0.91**	**0.95**	**0.33**	-	-	**85.42**

Table 2.6: Our results for Schwarz et al. dataset.

Features	SVM cost factor	F1	F1 (reliable class)	F1 (non reliable class)	Weighted Accuracy (%) $\lambda = 1$	$\lambda = 2$	$\lambda = 3$
Links	1	0.73	0.85	0	73.15	-	-
	2	0.73	0.85	0	-	57.66	-
	3	0.46	0.39	0.28	-	-	50.39
Links + Commercial	1	0.73	0.85	0	73.15	-	-
	2	0.73	0.84	0	-	57.63	-
	3	0.3	0.12	0.41	-	-	51.74
Words (removing stopwords)	1	0.74	0.85	0.14	73.86	-	-
	2	0.68	0.79	0.38	-	61.57	-
	3	0.55	0.63	0.44	-	-	58.65
Words (keeping stopwords)	1	**0.75**	**0.85**	**0.24**	**74.63**	-	-
	2	**0.69**	**0.79**	**0.41**	-	**62.93**	-
	3	**0.59**	**0.68**	**0.45**	-	-	**59.81**
All (removing stopwords)	1	0.74	0.85	0.15	73.88	-	-
	2	0.68	0.79	0.38	-	61.58	-
	3	0.55	0.62	0.44	-	-	58.39
All (keeping stopwords)	1	0.75	0.85	0.24	74.53	-	-
	2	0.7	0.79	0.4	-	62.89	-
	3	0.59	0.67	0.45	-	-	59.72

Table 2.7: Our results for CLEF eHealth dataset.

With **cost factor 3**, the detection of the minority class is slightly improved. As for the combination of features, both the word-based and the combination of all features (maintaining the *stopwords*) offer the same performance, but the former was selected because it generates a simpler model.

2.1.4.4 CLEF eHealth results

This was the largest dataset in our experiments, and it also presents an imbalance problem between classes. In contrast with Schwarz et al., a stratified **5-fold cross validation** could be applied given the larger number of data points. The obtained results are shown in Table 2.7.

For all cost factor values, the word-based model that maintains the *stopwords* was the one that offered the best results, yielding a reasonable minority or non-reliable class detection.

2.1.5 Conclusions of the additional experiments

Each of these additional datasets was different both in terms of content and task. Moreover, the original collection by Sondhi and colleagues was fully balanced, while the new datasets were clearly imbalanced. Nevertheless, some interesting conclusions can be drawn from these experiments.

The obtained results **reinforce** the main insights of the original study. In all experiments the best strategies were those that exploit bag-of-words features or merge all feature sets. The empirical evidence additionally suggest that keeping stopwords leads to enhanced performance.

2.1.6 Participation in the TREC 2020 Health Misinformation (HM) Track

As an applied use case, we tested this classical classification technology in the context of the TREC Health Misinformation Track. In 2020, the track focused on misinformation related to COVID-19 and SARS-CoV-2. Our understanding of this disease was constantly evolving and keeping track of objective and high quality information was critical. A solid retrieval system should be able to return scientifically accurate documents.

In this subsection, we explain the characteristics of the runs submitted by our team, **CiTIUS**, for the TREC 2020 Health Misinformation Track, and discuss our results. Our runs in 2020 represented an exploratory approach to leverage existing labelled data to build a classic reliability classifier and to test it with TREC Health Misinformation data.

2.1.6.1 Documents and topics

In the TREC 2020 Health Misinformation Track, a news corpus from January 2020 to April 2020 was provided. The documents were obtained from CommonCrawl News, which contains news articles from all over the world.

Topics attempt to model how people search for health advice online. Fifty topics with a fixed structure were provided. All topics have a numerical identifier, title, description, answer, evidence, and narrative, as it can be seen in Figure 2.2. The title field has the form of a pair of treatment and disease, where the disease is always COVID-19. The description is formulated

```
<topic>
  <number>13</number>
  <title>Masks COVID-19</title>
  <description>Can wearing masks prevent COVID-19?</description>
  <answer>yes</answer>
  <evidence>https://www.who.int/emergencies/diseases/novel-
  coronavirus-2019/advice-for-public/when-and-how-to-use-
  masks</evidence>
  <narrative>The widespread wearing of masks may be crucial in
  reducing the rate of transmission of COVID-19. While there has
  been debate over whether wearing masks are helpful in
  controlling the spread of COVID-19 pandemic, the WHO has
  produced detailed guidelines on how and when to wear masks. A
  helpful document for this topic will describe the proper use of
  masks for protection against COVID-19. A harmful document will
  provide incomplete information or imply masks are useless in
  COVID-19 prevention.</narrative>
</topic>
```

Figure 2.2: A TREC 2020 Health Misinformation Track topic (Topic 13).

as a question, which contains treatment, effect, and disease. The answer corresponds to the medical consensus at the time of topic creation. Finally, the remaining fields were not intended to be used by the systems, but only by human assessors to produce the assessments (*qrels*).

2.1.6.2 Retrieval baseline

For indexing and processing the collection, we considered different state-of-the-art tools, such as Terrier [124] or Lucene [115]. However, we decided to use **Anserini** [174], which is Lucene-based, but offers practical advantages to support the needs of this track.

For all the runs, the title field was used to produce the search query. We decided to use a bag-of-words approach, where at least one term or clause must match for a document to appear in the results. We selected a classical **BM25** [147] approach, setting the length normalisation parameter (b) to 0.75 and the TF weight upper-bound parameter (k_1) to 1.2. The first retrieval baseline was generated using Pyserini[12], Anserini's Python implementation. This facilitated the integration with the rest of the elements in our technology (our reliability classifier is also developed in Python).

The baseline was combined with other techniques, such as BERT sentence-similarity or our classic reliability classifier, in order to produce a final estimation of the presence of misinformation.

2.1.6.3 Classic Reliability classifier

Given a retrieved webpage, it was passed to classifiers built from the three collections described above (Sondhi's, Schwarz's and CLEF eHealth) and their predictions were aggregated to produce a final estimation. To that end, we built a model from each training collection using the best combination of features, as described above.

[12] https://github.com/castorini/pyserini

Given a document, Equation 2.2 shows its reliability score, where *pred_CLEF*, *pred_Sondhi* and *pred_Schwarz* are each model's prediction for the test document and the weights were set to the relative size of these three training collections:

$$Reliability(doc) = 0.97 \cdot pred\_CLEF + 0.027 \cdot pred\_Sondhi + 0.006 \cdot pred\_Schwarz \tag{2.2}$$

2.1.6.4 Submitted runs

Total Recall Task. In this task, the main goal was to retrieve documents that promulgate misinformation. To that end, documents contradicting the topic's answers were assumed to be **misinformation**. We submitted three different runs or variants to this problem.

The first one (**CiTIUSCrdTot**) applies first the BM25 retrieval baseline described before. Next, we ranked the *n* retrieved documents based on our reliability classifier's score (ranked by increasing reliability). We only kept the top ten thousand non-reliable documents in this ranking. We are aware of this being a **naive method**. For example, it ignores the matching between the description field and the retrieved pages, and just estimates misinformation based on the reliability of **the entire page**. In any case, we thought it was a natural baseline against which more sophisticated baselines could be tested.

The second run (**CiTIUSCrdRelTot**) applied a voting method, Borda Count [109], to combine two rankings: the original relevance ranking (BM25) and the reliability-based ranking.

The last run (**CiTIUSSimTot**) was the most sophisticated variant. A **hand-crafted expression** was created for each topic by combining description and answer fields. An example could be *Vitamin D cures COVID-19*, since we are looking to promulgate misinformation. Given the initial BM25 ranking (with queries produced from the titles of the TREC topics), we ranked the *n* retrieved documents based on maximum sentence similarity between the hand-crafted expressions and all sentences in each document (where sentences were represented using BERT). To this aim, we used Sentence Transformers[13] Python library, which offers several pre-trained models for embeddings generation, and then we applied cosine similarity between sentences.

AdHoc Retrieval Task. Unlike the previous task, here the main goal was to recover **correct information**. To that end, sites supporting the correct topic's answers were assumed to be relevant. We submitted four different runs or solutions to this problem.

The first one (**CiTIUSCrdAdh**) applied the BM25 retrieval baseline described before. After that, we ranked the *n* retrieved documents based on our reliability classifier's score but, in this case, we promoted highly reliable sites (the top thousand documents submitted were a ranking of documents by decreasing reliability).

[13] https://github.com/UKPLab/sentence-transformers

Runs	R-Precision
CiTIUSCrdTot	0.0105
CiTIUSCrdRelTot	**0.0354**
CiTIUSSimTot	0.0332
Median	0.0976

Table 2.8: Our results for the Total Recall Task.

Runs	CAM_MAP (us, co, cr)	NDCG (us, co, cr)	Comp. (harmful-only)	Comp. (helpful-only)
CiTIUSCrdAdh	0.0037	0.0412	0.0082	0.0586
CiTIUSCrdRelAdh	0.0355	0.1393	0.0475	0.1721
CiTIUSSimAdh	0.0252	0.1212	0.0351	0.1207
CiTIUSSimRelAdh	**0.0793**	**0.2353**	**0.0600**	**0.2376**
Median	0.1389	0.3308	0.0747	0.337

Table 2.9: Our results for the AdHoc Retrieval Task.

The second run (**CiTIUSCrdRelAdh**) applied a voting method, Borda Count [109], to combine relevance and reliability, and kept the top ranked documents.

The third run (**CiTIUSSimAdh**) consisted of producing a **hand-crafted expression** for each topic by combining description and answer fields. An example could be *Vitamin D does not cure COVID-19* (now, we are looking to promulgate correct and relevant information). After obtaining the title-based BM25 baseline, we ranked the n retrieved documents based on maximum sentence similarity between the new hand-crafted expression and all sentences in each document. As in the previous task, we used the Sentence Transformers library and cosine similarity.

Finally, the last solution (**CiTIUSSimRelAdh**) applied a sentence-similarity strategy again. However, it also used Borda Count to combine both rankings, relevance and similarity.

2.1.6.5 Results

Total Recall Task. The R-Precision results for the total recall task are shown in Table 2.8. All our methods performed worse than the median performance of the participants in the task. The classifier-based strategy (**CiTIUSCrdTot**) was the worst performer. It appears that this word-based document-level classification is too rough (and perhaps biased towards the topical words used in the training data). It must also be noted that the estimation of relevance combined with the reliability classifier (**CiTIUSCrdRelTot**) yields to better performance than that achieved by the reliability classifier alone. This suggests that relevance estimation should be kept as an integral part of the system. The BERT-based approach (**CiTIUSSimTot**) worked better than the classifier-based strategy but we did not combine it with any relevance information (because we could only submit three official runs). This suggests that the combination of **CiTIUSSimTot** with relevance information might lead to further benefits in terms of performance.

AdHoc Retrieval Task. This task was focused on obtaining credible and correct information. To that end, the assessments were created based on the concepts of *usefulness*, *correctness*, and

credibility.

The organisers designed specific measures to account for these aspects (e.g. compatibility, that measures the similarity to an ideal ranking) [43, 42]. However, they also evaluated runs in terms of traditional relevance measures (e.g. NDCG). Our results are shown in Table 2.9.

Again, our basic strategies fared worse than the median participant. The **CiTIUSSimRelAdh** run, which combined BERT-based similarity with the relevance ranking, produced our best results. The classifier-based variant was our worst performer.

2.1.6.6 Lessons learned from TREC

The document-level reliability classifier, presented in Section 2.1, generalised poorly when applied to data from the TREC HM 2020 challenge. In any case, our participation in TREC represented a major milestone and a valuable learning experience. This happened at the beginning of this doctoral research and the main goal was not to pursue state-of-the-art performance, but rather to put in place the necessary building blocks for a search system able to search for misinformation.

As reported above, we additionally evaluated a naive sentence similarity solution based on BERT. This solution seems to perform better, but it was still too simple. Our experiments also highlighted the necessity of effectively combining evidence of relevance and reliability for the accurate detection of misinformation.

2.1.7 Final remarks

In this section, a replicability study of classical reliability detection was presented. The main objective was to reproduce the experiments performed by a reputed research team and try to confirm the conclusions extracted from the original study. Our findings have served to emphasize that word-based models, or those that combine all features, are the most promising classic alternatives for discerning reliable from unreliable websites.

We have also tested this predictive technology against two further and highly different datasets and the conclusions remained the same. This gives us the confidence to state that the research presented in the original paper establishes a good classical reference for reliability detection in online data.

Finally, as a further test of transferability, this algorithm was exploited by our team in the TREC 2020 Health Misinformation Track[14], which tackles misinformation about COVID-19 and its treatments. In order to replicate the experiments presented in this work, the code is available for the research community at Github[15].

[14] https://trec-health-misinfo.github.io
[15] https://github.com/MarcosFP97/Health-Rel

	Trustworthiness (T)	Readability (R)	Useful (T&R)
# Positive	10,405	3,102	1,567
% Positive	73%	20%	12%
# Negative	3,820	12,455	11,488
% Negative	27%	80%	88%

Table 2.10: Label distribution in the CLEF eHealth dataset.

This initial exploratory research motivated us to further study how unreliable information is transmitted in the Web and how it is perceived by users. It also incited us to further analyse the effect of combining different features and, additionally, to consider the incorporation of new models using **BERT** [50]. This language modelling approach, which extracts a contextual representation of words, has been proven to be successful in the field of Natural Language Processing (NLP).

2.2 Comparing Traditional and Neural Approaches for detecting Health-related Misinformation

In this section, we evaluate the performance of traditional classification approaches, such as SVMs or KNNs, and newer BERT-based models for detecting health-related misinformation. To that end, we employed the CLEF 2018 Consumer Health Search task dataset. This task focuses on providing high-quality health-related search results to non-expert users. Different experiments were performed using target variables such as trustworthiness, readability, and the combination of both. Following Hahnel et al. [76], we consider that for a document to be useful it should not only be trustful but also understandable by non-expert users.

The main objective is to provide a thorough comparison between recent deep Natural Language Processing (NLP) models and traditional algorithms for the identification of poor quality online contents (untrustworthy and difficult to read web pages). We pay special attention to the behaviour of the models under realistic conditions (low training data). To that end, our study includes a report on the influence of the amount of training data in the effectiveness and the training time of the different models.

2.2.1 Dataset

To perform this comparison, we selected the CLEF 2018 Consumer Health Search task dataset [86], which focuses on the effectiveness of health-related information provided by search engines. The search task aims at helping non-expert users who are looking for health-advice. The dataset contains webpages obtained from CommonCrawl[16]. The creators of the dataset defined an initial list of potentially interesting sites and then, they submitted queries against a search

[16] http://commoncrawl.org/

engine to retrieve the final URLs. The initial list was manually extended by adding sites known to be either trustful or untrustful.

The assessments were provided by human assessors from Amazon Mechanical Turk. The turkers labelled the documents with respect to three different query-dependent dimensions: relevance, trustworthiness, and readability. In our experiments we consider only the last two variables.

Both dimensions of interest were judged on an eleven point scale, from 0 to 10. In our case, we wanted to approach the problem as a two-class classification challenge and, thus, we converted the original scores into binary variables. To that end, we removed the middle values (from 4 to 6) and mapped the extreme values to trustful/untrustful and readable/non-readable respectively. Table 2.10 reports the main statistics of the resulting datasets. We also tested classifiers for the task of distinguishing between *useful* documents for non-expert end users (i.e., trustworthy and readable) and *non-useful* documents (the remaining documents). With this goal in mind, we labelled useful documents as those that are both trustworthy and readable (third column in the table).

2.2.2 Experimental Design

We employed a 5-fold stratified cross-validation strategy in all the experiments. To address the imbalance in data labels, we also applied a cost-factor strategy [78, 110] in those learning methods whose implementation supports it[17]. We decided to set this cost-factor to the proportion between the classes for each experiment.

All experiments were conducted using the same docker container environment, an image with Ubuntu 18.04 and Python 3.7.3 version. The host machine also had 32GB of RAM, 240GB of storage, an Intel(R) Core(TM) i7-9750H CPU @ 1.60GHz, and a Nvidia Tesla V100S 32GB GPU, which was suitable for the BERT experiments.

2.2.2.1 Traditional models

We employed two variants for these experiments. The first consisted of a model where each word in a document was considered as a different feature, weighted by its normalised frequency. The second was equivalent, but stopwords were removed. The vocabulary was pruned to only consider terms present in at least 10% of the training corpus in both variants. We also applied a standardisation of the features (to get 0 mean and 1 standard deviation).

[17] We employed https://scikit-learn.org/stable/ (version 0.24.1)

- **SVM**. Following [155], we used a support vector machine implemented as part of the SVMlight toolkit[18] [87].

- **Random Forest (RF)**. We used Random Forest scikit-learn default implementation (100 trees were used and the Gini index was the criterion to measure the quality of a split).

- **Naive Bayes (NB)**. We used Naive Bayes scikit-learn default implementation, utilising the Multinomial Bayes variant, which is particularly recommended for imbalanced data problems.

- **KNN**. We used scikit-learn's default implementation of the KNN classifier ($k = 5$ neighbours).

For the models whose implementation supports cost weighting (SVM and RF) we also ran experiments with cost-weighting variants[19].

2.2.2.2 BERT-based models

For neural approaches, we considered BERT-based models [50]. These are pre-trained neural networks based on transformers, and have led to state-of-the-art solutions for many NLP tasks. We employed the **DistilBERT base** model (uncased version) [149] and **DistilRoBERTa base** model from the HuggingFace Transformers library [172]. The first has 6 layers, 768 hidden, 12 heads, and 66M parameters, while the second has the same number of layers, hidden and heads, but 82M parameters. These are light models obtained from larger ones, such as BERT base [50] or RoBERTa base [105]. The distilled models reduce the number of layers by a factor of 2, and the number of parameters by 40% while retaining 97% of the original performance [149].

These models were fine-tuned for our task in each fold. For the training process, 4 epochs and a 10% validation split were used, with a learning rate of 2^{-5}, a training batch size of 32, and a validation batch size of 64 instances.

BERT models have an input limit of 512 tokens. This was a challenge since the majority of the documents were larger. We trained the models with the first 512 tokens of each training document. At testing time, two different approaches were evaluated: i) making the prediction using only the first 512 tokens of the test document, or ii) segmenting each test document into 512-token chunks, passing the classifier on each chunk, and returning a final score that is the prediction score averaged over all chunks (aggregation strategy).

[18] Using default parameter setting (kernel linear and $C = [avg.\ x * x]^{-1}$). We employed the SVMlight Python wrapper with this configuration.

[19] Scikit-learn does not support cost-weighting for NB and KNN.

	Cost factor	F1 macro	F1 trustful	F1 untrustful
SVM (stopword removal)	1	0.57	0.84	0.3
SVM	1	0.57	0.83	0.31
SVM n-grams (stopword removal)	1	0.57	0.84	0.29
SVM n-grams	1	0.57	0.84	0.3
RF (stopword removal)	1	0.57	0.84	0.3
RF	1	0.57	0.84	0.29
Naive Bayes (stopword removal)	1	0.59	0.76	0.41
Naive Bayes	1	0.59	0.78	0.39
KNN (stopword removal)	1	0.6	0.8	0.39
KNN	1	0.59	0.82	0.36
DistilBERT	1	0.61	0.82	0.39
DistilRoBERTa	1	0.59	0.82	0.36
DistilBERT (aggregation)	1	0.58	0.83	0.33
DistilRoBERTa (aggregation)	1	0.61	0.84	0.38
SVM (stopword removal)	2.72	0.56	0.7	0.42
SVM	2.72	0.57	0.71	0.42
SVM n-grams (stopword removal)	2.72	0.57	0.71	0.43
SVM n-grams	2.72	0.57	0.71	0.43
RF (stopword removal)	2.72	0.57	0.84	0.29
RF	2.72	0.56	0.84	0.27
DistilBERT	2.72	0.6	0.74	0.45
DistilRoBERTa	2.72	0.59	0.72	0.46
DistilBERT (aggregation)	2.72	0.57	0.69	0.45
DistilRoBERTa (aggregation)	2.72	0.58	0.7	0.46

Table 2.11: Trustworthiness results obtained when setting or not the cost-factor to the proportion between classes.

2.2.3 Experimental Results

A set of experiments was performed for each target classification problem. We report the results for each of the different dimensions and models, providing the F1-score (harmonic mean between precision and recall) for each class and the macro average F1 (unweighted mean of F1-score per class).

2.2.3.1 Trustworthiness

The first dimension considered was trustworthiness. For this task, there is no substantial difference between the models (see Table 2.11). KNN and NB seem to be slightly superior to the other classic models and comparable to the best BERT-based variants.

With cost-weighting settings, the models tend to improve the detection of the minority class (untrustful), but the relative merits of the models remain essentially the same. Only RF shows here a distinctive behaviour, as its cost-weight variant decreases performance in terms of F1 untrustful.

Stopword removal had no substantial effect and the use of n-grams (bigrams and trigrams) did not bring any improvement (that is why it is only reported for SVMs). On the other hand, the aggregation strategy for BERT models did not yield any substantial advantage over a prediction that is solely based on the leading chunk. Making predictions with a single chunk of the

	Cost factor	F1 macro	F1 readable	F1 non-readable
SVM (stopword removal)	1	0.5	0.13	0.86
SVM	1	0.49	0.12	0.86
SVM n-grams (stopword removal)	1	0.49	0.11	0.86
SVM n-grams	1	0.49	0.12	0.86
RF (stopword removal)	1	0.51	0.16	0.86
RF	1	0.51	0.16	0.86
Naive Bayes (stopword removal)	1	0.59	0.33	0.84
Naive Bayes	1	0.59	0.33	0.84
KNN (stopword removal)	1	0.52	0.21	0.82
KNN	1	0.52	0.2	0.83
DistilBERT	1	0.5	0.19	0.81
DistilRoBERTa	1	0.49	0.16	0.81
DistilBERT (aggregation)	1	0.51	0.2	0.82
DistilRoBERTa (aggregation)	1	0.49	0.15	0.82
SVM (stopword removal)	4.02	0.51	0.3	0.72
SVM	4.02	0.5	0.31	0.68
SVM n-grams (stopword removal)	4.02	0.52	0.32	0.72
SVM n-grams	4.02	0.52	0.33	0.71
RF (stopword removal)	4.02	0.52	0.17	0.86
RF	4.02	0.53	0.18	0.87
DistilBERT	4.02	0.47	0.27	0.67
DistilRoBERTa	4.02	0.5	0.3	0.69
DistilBERT (aggregation)	4.02	0.49	0.28	0.7
DistilRoBERTa (aggregation)	4.02	0.48	0.27	0.69

Table 2.12: Readability results obtained when setting or not the cost-factor to the proportion between classes.

test document is computationally convenient, and our experiments suggest that this approach is comparable to a more thorough prediction based on the entire test document.

Overall, these results suggest that BERT models are unable to improve over simpler (and computationally less expensive) approaches. This could be related to the lack of large amounts of training data. In Section 2.2.3.4, we further analyse the models under varying training sizes.

2.2.3.2 Readability

In the readability experiments the objective was to detect the documents labelled as non-readable from the collection. The results in the readability experiments (see Table 2.12) show that the traditional algorithms perform better than BERT models. In particular, Naive Bayes achieves the best performance overall. When we set the $cost\text{-}factor = 4.02$ (notice that in this case the majority class was the non-readable), conclusions remain the same. Again, removing stopwords had no substantial effect on performance and the BERT-based models do not benefit from the aggregation approach.

These results suggest that determining readability can be effectively addressed with standard word-based technology. Even a simple bag-of-words model using a traditional learning method (like Naive Bayes or KNN) forms a solid classifier, comparable to the best neural models. One could argue that readability classification is essentially about distinguishing between the usage of simpler versus complex language. Our experiments show that such a goal can be competently tackled by classic NB technology.

	Cost factor	F1 macro	F1 useful docs	F1 non-useful docs
SVM (stopword removal)	1	0.51	0.1	0.92
SVM	1	0.5	0.07	0.93
SVM n-grams (stopword removal)	1	0.51	0.09	0.93
SVM n-grams	1	0.5	0.07	0.93
RF (stopword removal)	1	0.5	0.07	0.92
RF	1	0.5	0.06	0.93
Naive Bayes (stopword removal)	1	0.59	0.3	0.88
Naive Bayes	1	0.6	0.32	0.88
KNN (stopword removal)	1	0.54	0.16	0.92
KNN	1	0.53	0.15	0.91
DistilBERT	1	0.56	0.2	0.91
DistilRoBERTa	1	0.53	0.12	0.93
DistilBERT (aggregation)	1	0.56	0.2	0.91
DistilRoBERTa (aggregation)	1	0.54	0.16	0.91
SVM (stopword removal)	7.33	0.57	0.3	0.84
SVM	7.33	0.54	0.29	0.79
SVM n-grams (stopword removal)	7.33	0.58	0.31	0.84
SVM n-grams	7.33	0.55	0.3	0.8
RF (stopword removal)	7.33	0.51	0.1	0.92
RF	7.33	0.51	0.09	0.92
DistilBERT	7.33	0.57	0.29	0.84
DistilRoBERTa	7.33	0.5	0.27	0.73
DistilBERT (aggregation)	7.33	0.55	0.29	0.81
DistilRoBERTa (aggregation)	7.33	0.49	0.26	0.72

Table 2.13: Usefulness results obtained when setting or not the cost-factor to the proportion between classes.

2.2.3.3 Usefulness (trustworthiness & readability)

We also performed experiments combining readability and trustworthiness. To that end, we considered as *useful* documents the ones labelled as both trustful and readable. This seems reasonable since non-expert users look for trustworthy and understandable health-advice on the Web [76]. The remaining documents are regarded as *non-useful* documents (highly technical or untrustful).

The results (see Table 2.13) suggest that, as was the case in the trustworthiness experiments, there is no substantial difference between traditional and BERT models. Only a slight improvement of Naive Bayes over the rest was found. Again, applying a cost-sensitive learning strategy, improves the minority class detection, but RF does not benefit from this technique.

2.2.3.4 Influence of the training set size

In order to evaluate the influence of the training set size on effectiveness and efficiency, we report here two experiments: one for trustworthiness and another one for readability.

We selected **Naive Bayes**, **KNN**, and **DistilBERT base** (keeping stopwords and without any cost-factor), which were the best performing models in the experiments reported above. A 5-fold cross-validation strategy was applied again, but in this case models were only trained using a percentage of the training fold (always ensuring a stratified sample). We considered 1%, 5%, 10%, 30%, 50%, 70%, and 100% of the available data.

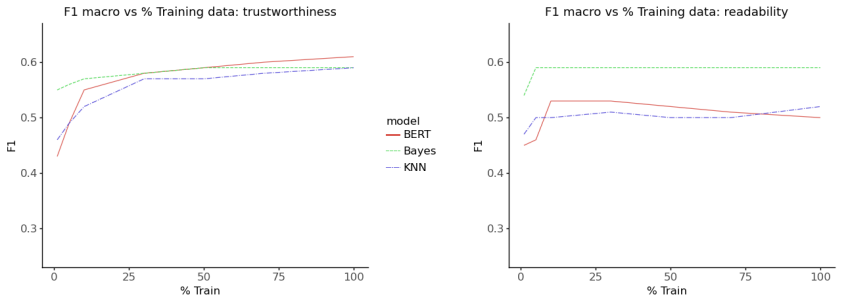


Figure 2.3: Variation of the F1 macro precision with percent training data used in trustworthiness and readability tasks.

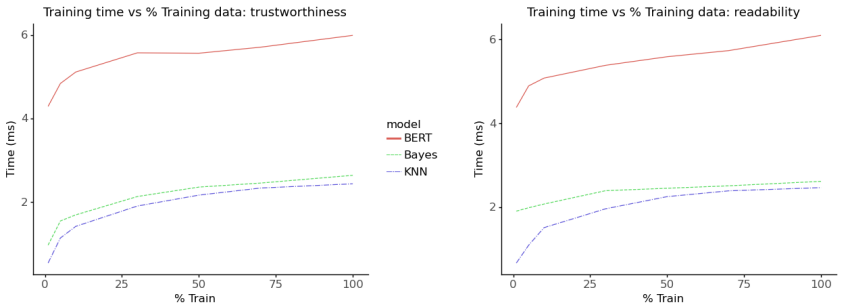


Figure 2.4: Variation of the training time (ms) with percent training data used in trustworthiness and readability tasks. Y axis in log scale.

In Figure 2.3, we depict how the F1 macro-precision of each model evolves with varying training data sizes. For trustworthiness (graph on the left), Naive Bayes clearly outperforms DistilBERT and KNN when training data is scarce. However, as we inject more training data, the performance of NB flattens, while the other models tend to benefit from the availability of more training examples. With the full training set, the three models perform roughly the same but the graph suggests that KNN and DistilBERT would keep improving and eventually beat the NB classifier.

For readability (graph on the right), Naive Bayes is the best performer over all training sizes. However, for all models, a low proportion of training examples seems to be sufficient. Observe that the performance of the three models tends to flatten (or even gets worse) with more than 20% of the training examples.

In Figure 2.4, we report the training times required by each model against the percentage of the training data. In both tasks, the training time taken by DistilBERT is much longer than that taken by the other models (we had to use a logarithmic scale for the representation). KNN is faster than Naive Bayes since it is a *lazy* approach (in training time it only stores the examples and learns no model).

Finally, we also computed the prediction time (time needed to classify a test instance). On average, Naive Bayes took 4.9 µs to predict, KNN 300 µs, and DistilBERT 0.002 µs. These results make sense since KNN has higher computational load in prediction time (needs to search for the neighbours). The DistilBERT model shows a surprisingly low average time, which could

be due to the fact that the underlying library is very optimized and takes advantage of the host GPU, while traditional models are only set to be executed in CPU.

2.2.4 Final remarks

The main lesson extracted from this comparative study is that, for these tasks and dataset, the added complexity of a neural model does not seem to be worthwhile. Sophisticated neural models were outperformed here by traditional models and the advantage of these classic methods was even more apparent with small training sets.

The results are modest overall and there is still room for improvement, as the tasks are difficult and more research effort is required. The main conclusion is that a traditional model such as NB is consistent (with very different sizes of training data), computationally efficient and should not be discarded (particularly considering that in many environments we have little training data).

This study opens up new lines of research related to how to detect health-related misinformation on the Web. A natural next step could be testing other strategies to deal with BERT input limit, such as generating summaries of the test documents and, subsequently predicting based on the summaries or, alternatively, using neural models that have no input limit, such as LongFormer [15].

Finally, we could also consider extending these experiments with BERT models already fine tuned for a document classification task.

3 A Multistage Retrieval System for Health-related Misinformation Detection

The previous chapter approached reliability estimation as a document-level problem and experimented with seminal methods that make a global analysis of the documents. However, combining multiple types of evidence (e.g., retrieval-based scores, supervised estimates or non-supervised estimates) is essential in searching for reliable documents. While a number of isolated studies have applied different features or signals for health-related reliability estimation, a complete picture of their effectiveness is still lacking. In this chapter, we try to fill this gap by constructing and evaluating a flexible multistage retrieval system able to incorporate and fuse multiple retrieval and classification stages. More specifically, we propose and evaluate multiple input signals from different sources and a number of combination methods that help to discern reliable from unreliable health information posted online.

Text-based features can play a major role in reliability estimation. Here, we want to further explore the ability of language-based features to enhance health-related misinformation detection and we have designed a complete series of experiments to evaluate them. This includes estimates at document and passage level, supervised and non-supervised methods powered by Deep Learning technology, re-ranking stages and different forms of fusion.

To evaluate our technological solutions we utilise the TREC 2020 Health Misinformation Track [42] as the main experimental framework. More specifically, the *total recall* task, whose goal is to identify all the documents conveying incorrect information for a specific set of topics, and the *ad-hoc retrieval* task, whose goal is to rank credible and correct information over incorrect information. Our flexible and modular technological solution can be easily adapted to support experiments for both search tasks.

The contents of this chapter were extracted from the following publications:

Fernández-Pichel, M.[a], Losada, D.E.[a], and Pichel, J. C.[a]. (2022). *A multistage retrieval system for health-related misinformation detection*. Engineering Applications of Artificial Intelligence, 115, 105211. The publication is available at: `https://doi.org/10.1016/j.engappai.2022.105211`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], Pichel, J. C.[a], and Gamallo, P.[a] (2021). *CiTIUS at the TREC 2021 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec30/papers/CiTIUS-HM.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], and Pichel, J. C.[a] (2022). *CiTIUS at the TREC 2022 Health Misinformation Track*. In: TREC proceedings. NIST. The publication is available at: `https://trec.nist.gov/pubs/trec31/papers/CiTIUS.H.pdf`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

3.1 Background

3.1.1 Combining multiple signals

Many previous studies in different areas have addressed the challenge of combining multiple pieces of evidence for a wide range of search or classification tasks. For example, Chenlo et al. [34] studied how to combine multiple signals for a blog distillation search task and other authors have defined and fused different features for computer vision [166, 160]. In our case, we focus on rank fusion techniques from the Information Retrieval (IR) field and we compare the performance of simple unsupervised rank fusion methods [64] (CombSUM and Borda Count [49, 9]) and more recent learning-to-rank (L2R) strategies [104] for the health misinformation detection task. Related to this, Benham and Culpepper [17] made a risk-reward analysis of multiple rank fusion methods. Their study was applied to a general retrieval task and, furthermore, the experimentation did not consider L2R methods. Our study focuses on a health-related search challenge, which naturally poses the need to incorporate not only retrieval-based features but also other types of AI-based signals (e.g. related to the credibility or correctness of the contents).

3.1.2 Systems for Health-related misinformation detection

Several teams have addressed the challenge of misinformation detection under the TREC experimental framework. Most have approached the problem with multiple complementary tools and have often utilised multiple re-ranking modules. More specifically, Pradeep and colleagues [136], from the University of Waterloo, proposed a multistage system that includes a final supervised re-ranker to promote reliability (based on a T5-3b model fine-tuned with external data). Bevendorff et al. [19] (Webis team) utilised the ChatNoir search engine [18] to obtain some baselines that are subsequently fed to a re-ranking module that re-organises the top results us-

ing certain query expressions. The best results were obtained by re-ranking the baselines with expressions generated from manual judgments that identified several relevant documents per topic (i.e., this approach requires explicit relevance feedback). On the other hand, Lima and colleagues [100] first performed a standard exact term matching retrieval, and then re-ordered the top documents in the ranking by fusing several signals. Here, we follow a similar approach but provide explicit evidence about the relative merits of the different modules. Moreover, we facilitate the use and extension of the platform, by making it publicly available.

3.2 Use case: Detecting COVID-19 misinformation

The TREC 2020 Health Misinformation Track[1] is oriented to search for misinformation in settings where the searcher knows the medical consensus at the time of issuing the query (for example, a social media moderator who wants to remove false health advice from the social media site, or a clinician who wants to alert about the increasing appearance of damaging recommendations).

To this end, the organisers provided a dataset, composed of news crawled from the web, and a set of topics. The collection was created from COVID-19 Common Crawl news extracted from January to April 2020. The topics represent health advice seeking requests. Each topic has a title, description or question, answer to the question, narrative and evidence field (see Figure 2.2). The description has the form of a question like "Can X Y COVID-19?", where X is a treatment and Y is one of the following effects: "cause", "prevent", "worsen", "cure", or "help". The answer field is "yes" or "no".

The track is divided into two subtasks: total recall and ad-hoc retrieval. The first subtask aims at identifying documents contradicting the topic's answer and, thus, the challenge is oriented to find documents conveying incorrect information. The second subtask focuses on promoting credible and correct information (documents that support the topic answer). Our methods were evaluated using both subtasks.

3.3 Input signals

A first contribution of our work consists of a novel architecture that incorporates multiple processing elements oriented to health misinformation detection. The architecture implements a pipeline that considers multiple pieces of evidence for ranking documents in terms of their estimated reliability to answer a given health-related information need. These input signals or features are computed over different stages.

The complete pipeline is shown in Figure 3.1. This system is freely available for the community to test and use[8]. Given a query, the process consists of four stages: one initial document

[1] https://trec-health-misinfo.github.io/2020.html

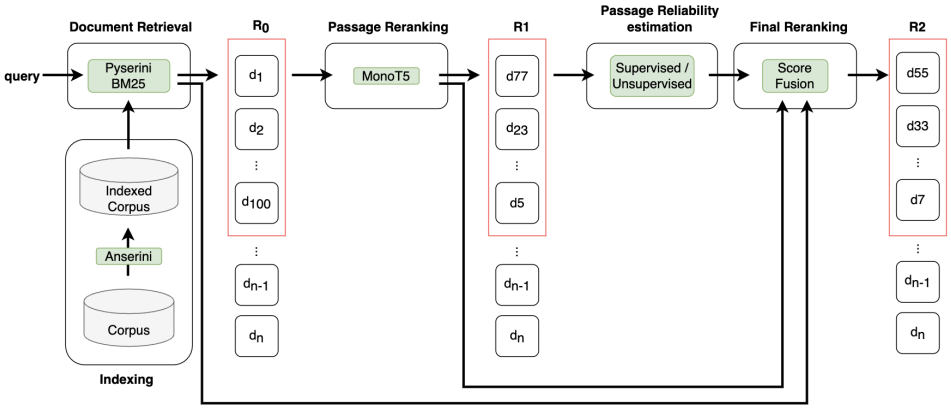


Figure 3.1: Full pipeline for health-related misinformation detection. After indexing the corpus, the system supports a document retrieval stage, passage-based re-ranking of the top retrieved documents, passage reliability estimation, and a final re-ranking stage that combines multiple signals.

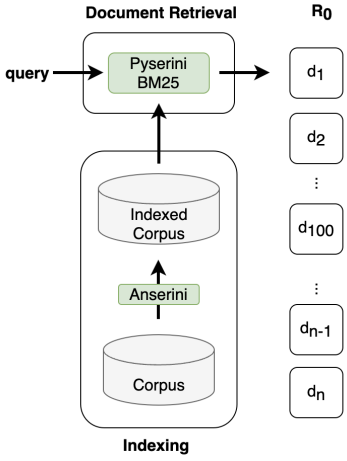


Figure 3.2: Document retrieval phase. The corpus is indexed with Anserini and, next, queries are executed against the resulting indexing. Search is done with the BM25 implementation from Pyserini.

retrieval phase that outputs an initial ranking of documents, a passage re-ranking phase that reorders the top 100 documents in the ranking according to the most relevant passages, a passage reliability estimation phase that implements either supervised or unsupervised techniques to obtain a score of how reliable/unreliable a passage is, and a final score fusion phase. The last fusion stage, which is discussed in more detail in the next section, accounts for the scores produced by all the elements of the pipeline in order to generate the final ranking (R2).

The features or input signals considered within this pipeline are:

– Document-level relevance:

In the first stage of the pipeline, the documents in the corpus are indexed using Anserini [175] (an open source textual corpus indexing engine). Given a health-related query, a *BM*25 [148] search for relevant documents is performed (see Figure 3.2). This outputs a ranking of documents ordered by decreasing estimated relevance.

*BM*25 is a well-known and effective IR model that does query-document matching based on standard IR weights (term frequency, inverse document frequency and document length normalisation). Equation 3.1 presents the *BM*25 document relevance score, where $tf(q_i, D)$ is the term frequency of q_i in the document D, L_D is the number of tokens in document D, and L_{avg} is the average number of tokens per document in the collection, respectively. k_1 is a parameter that controls the term frequency saturation, while b is a parameter that tunes the effect of length normalisation.

The *IDF* component (Equation 3.2) is based on $df(q_i)$, which is q_i's document frequency in the collection, and N, the number of documents in the collection[2].

$$BM25(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{tf(q_i,D)}{tf(q_i,D) + k_1 \cdot \left(1 - b + b \cdot \frac{L_D}{L_{\text{avg}}}\right)} \tag{3.1}$$

$$IDF(q_i) = log\left(1 + \frac{N + df(q_i) + 0.5}{df(q_i) + 0.5}\right) \tag{3.2}$$

Over the years, multiple implementations of *BM*25 have been made available. Kamphius and colleagues showed recently [90] that there are no major differences between eight variants of *BM*25. We employed the Pyserini[3] library, which employs the Lucene implementation of *BM*25. The experiments were run with the following parameter setting: $k_1 = 0.9$ and $b = 0.4$ (values which are in the recommended range for these two parameters).

– Passage-level relevance:

In this second stage, we intend to skip the noisy content in each document and focus solely on the passage most similar to the query. Our approach is based on sequence-to-sequence models for document ranking as described in Nogueira et al. [122].

In NLP, with the emergence of the Transformer architecture [161], various transfer learning approaches pre-train a given model for a generic task, and then fine-tune it on specific downstream problems. Nogueira and his colleagues proposed using T5 [142], by Google, for document ranking. This architecture attempts to handle all downstream tasks into a text-to-text format. In contrast to BERT-like architectures [50], the text-to-text framework uses the same loss function and hyperparameters for all NLP tasks. Inputs to the model are encoded in such a way that the model identifies the task, and the output is always in the form of text.

T5 is pre-trained for a denoising task, masking a sequence of words from the sentence and training the model to predict these masked words (see Figure 3.3). This gives the model

[2]The constant 1 is added to avoid negative scores, which would otherwise occur when $df(q_i) > N/2$.
[3]`https://github.com/castorini/pyserini`

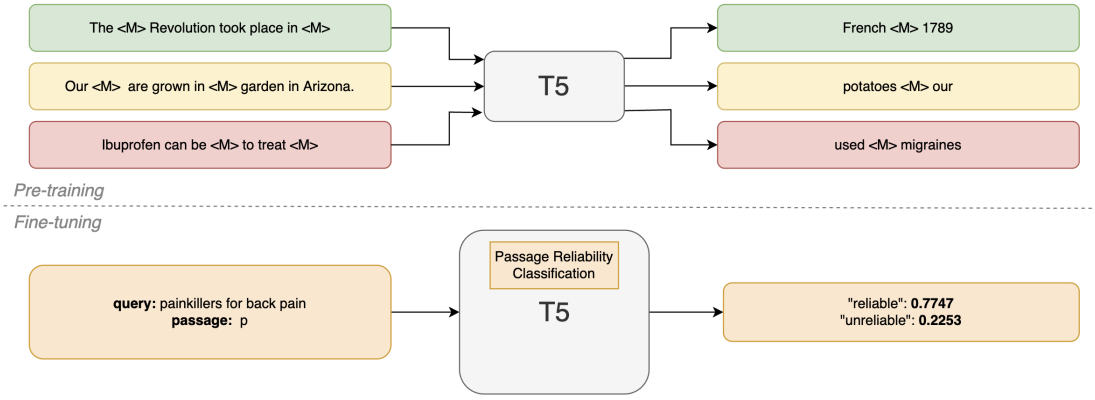


Figure 3.3: T5 fine-tuning for ranking passages (example in the upper part from Raffel et al.[142]). The pre-training stage tunes the model for general language understanding tasks and, next, the model is fine-tuned for the estimation of relevance at passage level.

the ability to learn general intricacies of the language. Afterwards, the model is fine-tuned on a downstream task with a supervised objective using the appropriate input. We employ this technology to classify a passage as relevant to a given query (see Figure 3.3). To that end, the query and document act as input sequences, and the model is fine-tuned to produce the tokens *"true"* or *"false"* depending on whether the document is relevant or not. At prediction time, probabilities for each token are computed using a softmax layer, which outputs the value used for ranking.

This model was fine-tuned with different datasets that are freely available at Pygaggle[4] library. In our case, we decided to experiment with the MonoT5 base model fine-tuned for passage re-ranking with Med-MARCO, a medical subset of the passage ranking dataset MS MARCO [121]. This data collection is oriented to relevance ranking for the biomedical domain.

The reliability of a document with respect to the query topic needs to be assessed based on query-related document's contents. To that end, the most relevant passage of each document is kept and a new ranking of documents is produced using passage-level relevance scores. Following standard practice, only the top documents from the initial ranking are re-ranked. We re-ordered the first 100 documents using their passage scores while the remaining documents (ranks greater than 100) were kept at their original positions[5] (see the output in Figure 3.4).

In order to determine the most relevant passage, a sliding window was applied to each document (see Figure 3.5). Following Pradeep et al. [137], we decided to set the window length to 6 sentences and its stride to 3 sentences. This is a reasonable setting, as passages of this length can contain a complete answer on a health-related topic. No op-

[4] `https://github.com/castorini/pygaggle`

[5] As a matter of fact, positions greater than 100, are actually fixed over the entire process (all remaining processing stages only act on the top 100 documents).

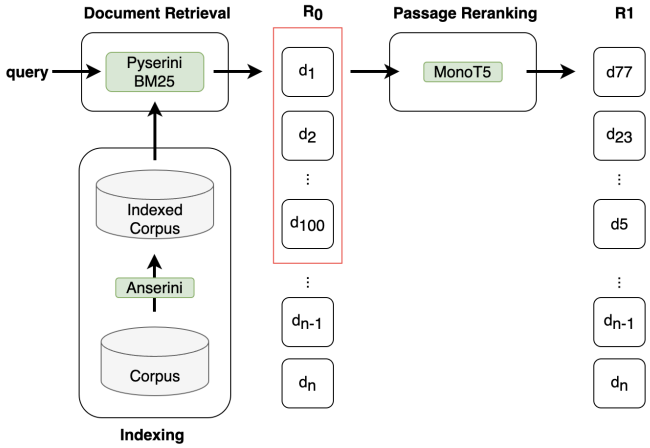


Figure 3.4: Passage re-ranking phase. The top retrieved documents from the initial ranking (R_0) are re-ordered based on the most relevant passages (passage relevance estimates obtained from MonoT5).

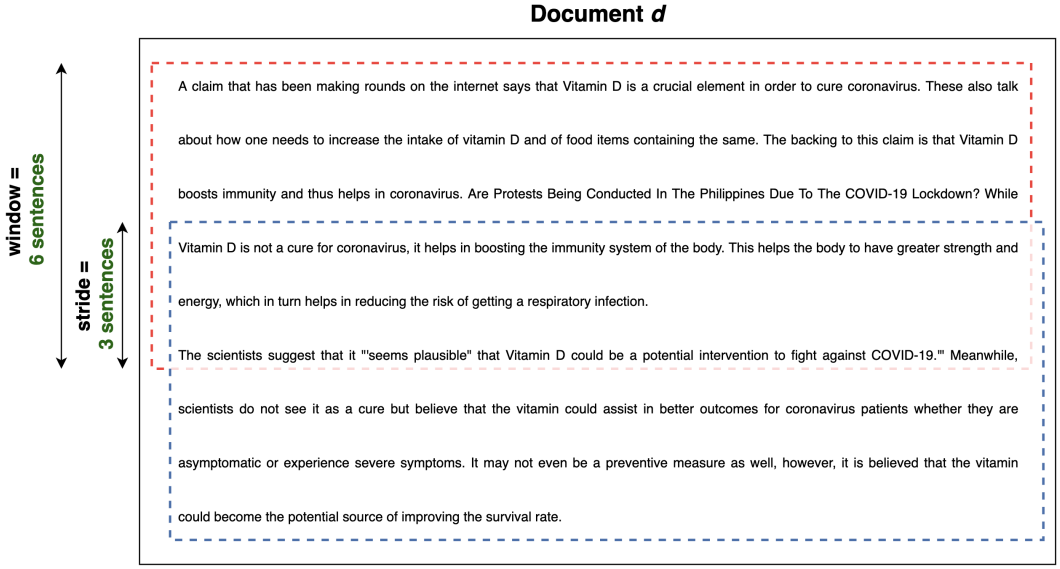


Figure 3.5: Sliding window ($window = 6$ and $stride = 3$ sentences) used to perform passage re-ranking.

timisation was made concerning these parameters. Each candidate passage was passed to the MonoT5 model described above and the passage yielding the highest score was selected to estimate the document's passage-level relevance.

– Passage-level reliability:

In this stage, the estimation of reliability of the extracted passage is considered as a new feature that might help in the misinformation identification process. We evaluated two alternative methods to predict reliability of the passages:

– Supervised: a T5 model was fine-tuned to classify a passage as reliable or unreliable with respect to a given query (see Figure 3.6). To this end, we fine-tuned the model with several training examples as detailed in Section 3.5.

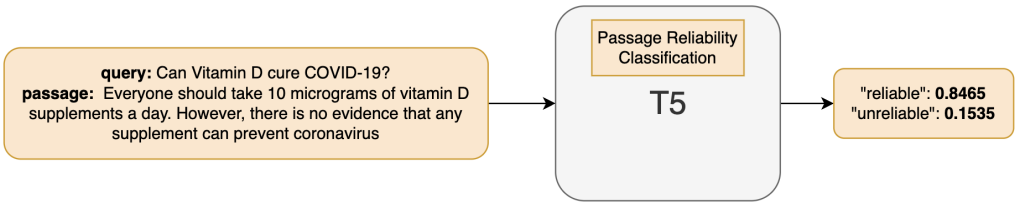


Figure 3.6: T5 fine-tuning process for passage reliability classification. The fine-tuning stage takes queries and passages labelled in terms of reliability.

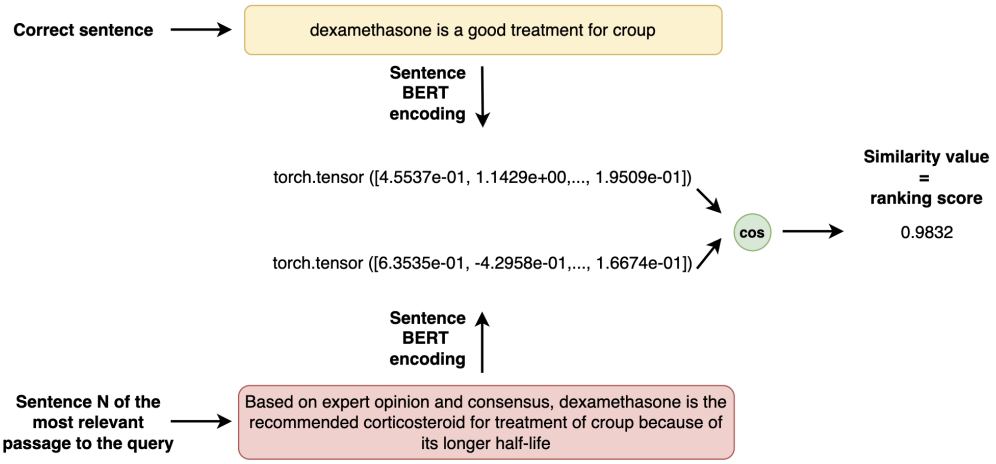


Figure 3.7: Unsupervised strategy for passage reliability detection. Sentences from the most relevant passages are represented with Sentence BERT and their similarity to the Sentence BERT representation of the query expression is computed.

The resulting classifier was run on the passages selected in the previous stage (passage-level relevance), obtaining a probability score for each passage (associated to the tokens *"reliable"* and *"unreliable"*). These reliability/unreliability probability values are used as input scores in the score fusion phase.

– Unsupervised: as an alternative "unsupervised"[6] strategy, we used a sentence similarity approach. We created hand-crafted true and false claims (depending on the subtask) from the given queries and we compared them with each sentence in the most relevant passage of the document (see Figure 3.7). Recall that the main use case of this search technology is oriented to users (e.g., moderators) who know the correctness/incorrectness of the claim and, thus, the truthfulness of the search topic is available and can be fed to the system.

  We encoded the input sentences with Sentence BERT models [144]. Previous studies have shown that these models perform much better than traditional BERT models for sentence similarity tasks [67]. The cosine similarity measure was applied on the obtained embeddings.

[6] We are aware that using the word "unsupervised" here might be consider an abuse of language, as these models are pre-trained with large amounts of data. However, in this context, by unsupervised, we mean that we do not apply an ad-hoc fine-tuning process to the original model.

3.4 Fusion strategies

To address the combination of document-relevance, passage-relevance and passage-reliability signals, two different approaches were compared: unsupervised rank fusion methods and learning-to-rank. These methods are described next. Observe that, in our setting (oriented to re-rank the top 100 documents), each top document always has the three input signals and, thus, the fusion consists of merging three ranked lists of size 100.

- Unsupervised Rank Fusion: This is a common technique in IR, which tries to respond to a user's information need by combining knowledge from the output of many retrieval systems [64]. Fox and Shaw proposed several unsupervised rank fusion methods (no training needed), named as the "Comb" family. CombSUM and CombMNZ, which are score-based, are the most effective methods. Other authors proposed combination strategies based on the ranking positions like Borda [49, 9]. In this work, we will evaluate the effectiveness of a score-based method (CombSUM) and a rank-based alternative (Borda):
  - CombSUM is a score-based technique that sums the scores that the document has in each ranked list. Equation 3.3 presents CombSUM, where L is the number of ranked lists to fuse and s_{lj} is the score for a concrete document j in a specific ranking l. In our case, the scores were first normalised[7].

$$score(d_j) = \sum_{l=1}^{L} s_{lj} \tag{3.3}$$

  - Borda Count is a rank-based technique that implements a voting scheme. Each document gets votes from each ranked list, and these votes are added. The number of votes depends on the document position in the list. Equation 3.4 presents Borda, where L is the number of ranked lists to fuse, n is the number of ranked elements, and p_{lj} is the position of document j in a ranking l.

$$score(d_j) = \sum_{l=1}^{L} n - p_{lj} + 1 \tag{3.4}$$

- Learning-to-rank (L2R) algorithms learn how to combine features extracted from query-document pairs through a training process [104]. There are three main classes of L2R methods: pointwise (predict the relevance degree of a single document) [38, 37, 39], pairwise (predict the preference between a pair of documents) [11, 26, 29], and listwise (predict the whole ranked list) [3, 30, 138]. In this work, we focus on the pointwise approach, which is the most common L2R method and requires fewer training examples compared with its L2R counterparts (see Section 3.5 for more details).

[7] The scores were normalised by dividing by the maximum value for each topic.

– Pointwise L2R. Given multiple features associated with each candidate document (three scores in our case), pointwise methods predict the relevance degree of each single document. To that end, some form of supervised learning is performed from training data. Given a split of training queries, we compute the three scores of each top ranked document (BM25 document relevance score, relevance score of the document passage that is the most similar to the query, and passage reliability score of the most similar passage), extract the reliability label of these documents from the ground truth judgements, and feed the 3-feature representations together with the target labels to a logistic regression classifier. This binary classification approach allows to learn how to combine the three predictors in order to estimate how reliable a retrieved document is. The resulting probability estimates are used to produce the final ranking:

$$P(Reliable|d_j) = \frac{1}{1 + e^{-c-\sum_{l=1}^{L} w_l \cdot s_{\mathrm{lj}}}} \tag{3.5}$$

where $P(Reliable|d_j)$ is the probability estimate of reliability for document j, L is the number of ranked lists to fuse (equal to three in our case), s_{lj} is the score of a document in a concrete ranking, and c and w_l are the parameters learnt by the logistic regression model from the training collection.

3.5 Experimental setup

The main contributions of this chapter can be summarised as follows:

– A complete pipeline for health misinformation detection is proposed. The two target tasks, total recall and ad-hoc retrieval, represent socially important scientific and technological challenges. Searching for unreliable information (total recall task) has a number of potential applications, including web content moderation or crawl filtering. Similarly, the ad-hoc-retrieval task is valuable to advance in solid search methods that promote correct and credible contents.

– The technology developed is freely available[8] and other researchers, practitioners and relevant stakeholders can reuse and adapt our technological solution. For example, it could be employed by moderators of a social media platform or a health-related website to identify and filter out unreliable contents. This contrasts with existing proposals that support health misinformation experiments, which often lack a full disclosure of their settings and do not inform about the relative merits of their relevant components.

[8] https://github.com/MarcosFP97/Multistage-Retrieval-System

– Different content-based features or information signals are introduced for estimating the occurrence of reliable/unreliable web contents. This includes search-based signals, at document and passage level, reliability estimators based on state-of-the-art deep learning models and fusion methods for combining evidence.

– A thorough analysis of performance is performed, including an ablation study, of the different signals and fusion methods. Our evaluation uses innovative metrics that consider relevance, harmfulness and helpfulness of the retrieved documents. More specifically, our study reports standard search metrics, such as those based on Average Precision, R-Precision and Normalised Discounted Cumulative Gain, and novel effectiveness measures, which estimate the overlap between the ranked output and two reference rankings of harmful and helpful documents.

– The empirical validation of the system provides interesting insights. For example, focusing the analysis on the most relevant passages stands out as a key component, which improves the retrieval of helpful documents and reduces the retrieval of harmful contents. On the other hand, passage reliability estimators are also beneficial but the limited availability of training data makes that the most solid estimates are derived from non-supervised methods. For combining evidence, our results suggest that simple score fusion techniques are superior to more advanced combinations based on learning to rank.

– We also analyse thoroughly the most effective variants in the light of the trade-off between the retrieval of helpful and harmful results and we demonstrate that our best performing approach attains competitive performance compared to the highly sophisticated systems submitted to TREC.

3.5.1 Total Recall

For this task, the notion of "relevance" is binary, and a relevant document is a document that provides incorrect information about the query topic. The official effectiveness metric is R-precision [47] (Equation 3.6). This evaluation measure computes, for each query, the precision at the R-th position of the ranking, where R is the number of relevant documents that the query has in the corpus:

$$R_{\text{prec}} = \frac{r}{R} \tag{3.6}$$

where r is the number of relevant documents found by the system at the top R positions. The reported R_{prec} figures are averages of the R_{prec} obtained over all available queries.

3.5.2 Ad-hoc Retrieval Task

The ad-hoc retrieval task aims at designing a retrieval system that promotes credible and correct information over incorrect information. Contrary to the previous task, this task considers a more sophisticated notion of relevance. There are multiple types of documents: useful and correct and credible, useful and correct and not credible, non-useful and incorrect, and so forth. Useful here means on-topic (i.e., a document that is topically relevant with respect to the query). Correctness refers to whether or not the document contains a definitive and correct answer to the topic question, while credibility refers to whether or not the document is considered credible by the assessor.

Given these three dimensions, a graded relevance scale was defined. The best documents (graded relevance=4) are those that are useful, correct and credible, while the worst documents (graded relevance=-2) are those that are useful, incorrect and credible. This last class of documents is really damaging because these documents seem useful –on-topic– and credible to the user but they provide incorrect information. Table 3.1 presents the full scale of grades of relevance. These grades of relevance were employed in a number of ways. First, some standard IR metrics can measure the quality of search results taking into account different levels of relevance. For example, the Normalized Discounted Cumulative Gain (*NDCG*) [47] is one of the official metrics utilised to evaluate the algorithmic solutions proposed for the ad-hoc retrieval task:

$$NDCG = \frac{DCG}{IDCG} \tag{3.7}$$

$$DCG = \sum_{p=1}^{n} \frac{2^{\text{rel}_p} - 1}{log_2(p+1)} \tag{3.8}$$

DCG, Discounted Cumulative Gain, defines the user's gain as a measure that grows as the user goes from top to bottom positions of the ranking. Under *NDCG*, the gain produced by each ranked document depends on its position. Gains from relevant documents at higher positions are greater than gains from relevant documents at lower positions. To that end, each gain is divided by a discounting factor ($log_2(p+1)$). The *DCG* values are normalised by dividing the *DCG* scores by the ideal *DCG* (*IDCG*, which represents the gains obtained by a perfect system that ranks documents by decreasing order of their actual relevance). *NDCG* can be computed at any cutoff but we report here the *NDCG* scores associated to the entire ranking (whose size is n). Observe that, under *NDCG*, rel_p scores cannot be negative. Following standard practice in the TREC Health Misinformation track, the computation of *NDCG* assigns 0-gain to all documents whose relevance degree is less or equal to 0.

Another IR metric considered in our study is the Convex Aggregation Measure of the Mean Average Precision (*CMAP*) [101]. *CMAP* combines the Mean Average Precision (*MAP*) [47] of usefulness, correctness and credibility as follows:

$$CMAP = \lambda_1 \cdot MAP_u + \lambda_2 \cdot MAP_{co} + \lambda_3 \cdot MAP_{cr} \tag{3.9}$$

where a uniform combination leads to $\lambda_1, \lambda_2, \lambda_3 = 1/3$. Each *MAP* score comes from inspecting the ranking with a different notion of "relevance": usefulness (MAP_u), correctness (MAP_{co}) and credibility (MAP_{cr}). The *MAP* score is the mean of the average precision (*AP*) values associated to multiple queries:

$$MAP_x = \frac{\sum_{i=1}^{|Q|} AP_x(q_i)}{|Q|} \tag{3.10}$$

$$AP_x = \frac{1}{r_x} \cdot \sum_{p=1}^{r_x} P(p) \cdot rel(p) \tag{3.11}$$

where x is u, co or cr. *AP* represents the area under the precision-recall curve. r_x is the number of relevant documents (number of useful, correct or credible documents, respectively), $P(p)$ is the precision at a cutoff p, and $rel(p)$ equals 1 if the item at rank p is a relevant document, and 0 otherwise. Observe that *AP* works with binary relevance values. The usefulness labels (third column in Table 3.1) are already binary and, thus, MAP_u can be straightforwardly computed. For computing the *AP* of correctness (MAP_{co}), documents that give no answer or documents that have not been assessed for correctness are assigned a score equal to 0 (i.e. 2 and -1 are transformed into 0). For computing the *AP* of credibility (MAP_{cr}), documents that have not been assessed for credibility are assigned a score equal to 0 (i.e. -1 are transformed into 0).

The graded relevance values were also employed to compute other innovative metrics, such as compatibility [43]. Compatibility estimates the similarity between a ranked list provided by an automatic system and an ideal ranking. Clarke and colleagues utilised Rank Biased Overlap (*RBO*) [170] to compute compatibility between an ideal ranking I and an actual ranking L as follows:

$$RBO(L,I) = (1 - pat) \cdot \sum_{p=1}^{\infty} pat^{\text{p-1}} \frac{|I_{1:\text{p}} \cap L_{1:\text{p}}|}{p} \tag{3.12}$$

where $I_{1:\text{p}}$ and $L_{1:\text{p}}$ represent the top p documents in I and L, respectively. The overlap between both rankings at the cutoff p is defined as the size of the intersection of these lists. *RBO* is then a weighted average across cutoffs from 1 to ∞, and $pat \in (0,1)$ models searcher patience.

Following [42], we calculate: i) compatibility helpful, where the ideal ranking is composed only of the documents whose relevance level is greater than zero (ordered by decreasing graded

Relevance Degree	Description	Usefulness	Correctness	Credibility
4	Useful, correct, credible	1	1	1
3	Useful, correct, not credible or no credibility judgement	1	1	0 or -1
2	Useful, no answer or no judgement for answer, credible	1	2 or -1	1
1	Useful, no answer or no judgement for answer, not credible or no judgement	1	2 or -1	0 or -1
0	Not useful, ignore answer and credibility	0	-	-
-1	Useful, incorrect, not credible or no judgement	1	0	0 or -1
-2	Useful, incorrect, credible	1	0	1

Table 3.1: Preference ordering for documents mapped to graded relevance. Usefulness=1 (0) means that the document is on-topic (off-topic). Correctness=1 (0) means that the document gives a correct (incorrect) answer to the health-related request. Correctness=2 means that the document gives no answer to the health-related request. Correctness=-1 means that the document was not manually judged in terms of correctness to the health-related request. Credibility=1 (0) means that the document was judged as credible (non-credible). Credibility=-1 means that the document was not judged in terms of credibility.

relevance), and ii) compatibility harmful, where the ideal ranking is composed only of the documents whose relevance level is negative (ordered by increasing graded relevance). A good system should score high on compatibility helpful and low on compatibility harmful. Additionally, a global compatibility score is reported as the difference between the compatibility helpful achieved by a system and its compatibility harmful.

3.5.3 Experimental details

The TREC 2020 Health Misinformation dataset contains 31 search topics that have both helpful and harmful documents[9]. Some of the signals described above require training data and, thus, we employed a three-fold splitting strategy. This means that the effectiveness results are averaged over the three test folds. In order to get comparable results for the non-supervised methods, these were also evaluated on the same test folds (and their results averaged). However, it is important to bear in mind that the unsupervised methods did not utilise any information from the training fold associated to each test fold.

To analyse the statistical significance of the performance difference between two systems or alternatives, we applied the Wilcoxon test on the paired values (one from each query). Parapar and colleagues [126, 127] recently compared several significance tests in the context of Information Retrieval experiments and showed that Wilcoxon test is a highly reliable test to compare retrieval systems (yields more statistical power and fewer type I errors). The reported results of

[9]Originally, the track organisers created 50 topics but, after building the relevance assessments, many of them ended up with only examples of helpful documents. We therefore focus on the 31 topics that have reliable and unreliable retrieval results.

statistical significance correspond with the entire set of 31 topics (each query result extracted from its corresponding test fold). The significance tests help us to determine whether or not each observed difference is anecdotal.

The following settings were utilised for the different parts of our architecture:

- **Document-level relevance (BM25)**: we used Pyserini's implementation[3] of BM25 with k_1 set to 0.9 and b set to 0.4.

- **Passage-level relevance (MonoT5)**: we used the Pygaggle[4] library and, more specifically, the MonoT5 model fine-tuned for passage re-ranking with Med-MARCO.

- **Reliability estimation (supervised methods)**: given the training queries available in the train fold, we fine-tuned a T5-base model with a constant learning rate of 3×10^{-4} for a variable number of iterations depending on fold size and with batches of size 8. We ran 2 training epochs and selected a maximum length of 512 tokens.

  We evaluated three strategies to fine-tune this classifier:

  - training with the (unmodified) query (description field) upfront + passage + label (`class q`).
  - training with a correct hand-crafted expression + passage + label (`class cs`).
  - training with an incorrect hand-crafted expression + passage + label (`class is`).

  The hand-crafted expressions were created from the description and answer fields of each topic. For example, for the question "Can Vitamin D cure COVID-19?", whose correct answer is "no", the two expressions were: "vitamin D can cure COVID-19" (incorrect hand-crafted expression) and "vitamin D can not cure COVID-19" (correct hand-crafted expression). The creation of these expressions from the description and answer field of each topic is inspired by previous studies on parsing [178].

  Given the labels assigned to the documents, we have opted to build a correctness classifier (by considering only the correctness label), a credibility classifier (by considering only the credibility label) or a correctness+credibility classifier (by considering the conjunction of both labels). After some preliminary experiments, we decided to adopt the latter as our reference reliability classifier.

- **Reliability estimation (unsupervised methods)**: the unsupervised strategy also involves creating hand-crafted expressions but, rather than using them to build a classifier, they are employed to search for similar sentences within each target passage. With this in mind, we

	Rprec Incorrect
Document relevance (DOC_REL)	0.1025
Passage relevance (PAS_REL)	0.1096

Table 3.2: Relevance-based search method results for the total recall task.

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
DOC_REL	0.2537	0.5101	0.1206	0.3186	0.1981
PAS_REL	**0.2871**$^{\uparrow}$	**0.5435**	**0.1180**	**0.3794**	**0.2614**

Table 3.3: Relevance-based search method results for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

used Sentence BERT library[10,11]. More specifically, we experimented with four different models:

- BERT Large fine-tuned only with NLI dataset [24] (`sim BERT-NLI`).
- BERT Large fine-tuned with NLI+STSB datasets [31] (`sim BERT-STSB`).
- RoBERTa Large fine-tuned with NLI dataset (`sim RoBERTa-NLI`).
- RoBERTa Large fine-tuned with NLI+STSB datasets (`sim RoBERTa-STSB`).

The final score consists of the average value of all the similarity scores computed between the hand-crafted expression and each passage sentence.

3.6 Results

3.6.1 Relevance-based search methods

First of all, we tested the performance of the stages of our pipeline that merely incorporate topic relevance, namely: the document relevance estimation phase (`DOC_REL`), and the passage relevance estimation phase (`PAS_REL`).

Results for both tasks are shown in Tables 3.2 and 3.3. The best method for both tasks is `PAS_REL`, which yields the best performance figures in all metrics. This strategy takes the ranking generated from document relevance scores and re-ranks the first 100 documents by decreasing passage relevance score. These results show that the passage-based strategy is effective. Scoring documents based on the most relevant passage leads to substantial improvements in performance and in one case the improvement is statistically significant. The passage relevance approach looks promising (and, on average, leads to higher effectiveness than that of document

[10] `https://www.sbert.net/`
[11] `https://pypi.org/project/sentence-transformers/0.4.1.2/`

relevance). However, the characteristics of this test set make it hard to reveal statistical significance. We analysed the individual (per-query) effectiveness scores and, for example, the PAS_REL variant leads to improved performance in 19 out of 31 queries (compatibility). With a larger query test we suspect that we could easily obtain improvements that are statistical significant. In any case, the improvements are not consistent across queries and, in the near future, we plan to further explore methods that incorporate query-dependent techniques (e.g., alternate between passage and document retrieval in a topic-dependent way).

Relevant passages represent a concise and on-topic representation of the document that eliminates content that is unrelated to the query. The relative merits of PAS_REL and DOC_REL clearly suggest that misinformation detection should concentrate on the most relevant extracts from the retrieved webpages.

We also ran some exploratory experiments where we combined the scores of document relevance and passage relevance. To that end, we employed the score fusion techniques described in Section 3.4. However, these tests did not result in any advance over PAS_REL alone and, thus, we adopted the passage relevance signal as the reference topic-relevance baseline for further experiments.

3.6.2 Reliability estimation at passage-level

Next, we evaluated the effectiveness of the passage reliability estimation methods. To this end, we re-ranked the top 100 documents by decreasing estimation of reliability of the most relevant passage. We experimented with the supervised and unsupervised reliability methods described in Section 3.3.

For the first subtask, total recall, these methods fail to outperform the relevance-based baseline (see Table 3.4). There are three methods whose performance is not statistically inferior to the baseline. However, no method yields performance figures higher than PAS_REL and, thus, the reliability signal alone is insufficient to find documents that include misinformation. Note that the best performing supervised alternative is PAS_RELIA_C_IS, which is the method that trains with the incorrect hand-crafted expression upfront. This is a natural outcome, as the total recall task aims at searching for incorrect documents.

For the second subtask, ad-hoc retrieval, the supervised methods again yield poor performance (see Table 3.5). In terms of compatibility harmful, the supervised strategies lead to substantial benefits but, for the remaining metrics (including global compatibility), performance is much worse than that of the baseline. This suggests that these methods have poor retrieval performance. The lack of on topic documents retrieved means that fewer are either helpful or harmful. The best supervised method is here PAS_RELIA_C_CS, which trains with the hand-crafted correct sentence. Again, this is a natural outcome, as the ad-hoc retrieval task aims to find reliable documents. On the other hand, the unsupervised methods seem to provide added

	Rprec Incorrect
Reference	
PAS_REL	0.1096
Supervised methods	
Passage reliability (class q) (PAS_RELIA_C_Q)	$0.0703^{\downarrow}$
Passage reliability (class cs) (PAS_RELIA_C_CS)	$0.0726^{\downarrow}$
Passage reliability (class is) (PAS_RELIA_C_IS)	0.0822
Unsupervised methods	
Passage reliability (sim BERT-NLI) (PAS_RELIA_S_BN)	$0.0826^{\downarrow}$
Passage reliability (sim BERT-STSB) (PAS_RELIA_S_BS)	0.0908
Passage reliability (sim RoBERTa-NLI) (PAS_RELIA_S_RN)	$0.0796^{\downarrow}$
Passage reliability (sim RoBERTa-STSB) (PAS_RELIA_S_RS)	0.0780

Table 3.4: Passage reliability estimation results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_REL	0.2871	0.5435	0.1180	0.3794	0.2614
Supervised methods					
PAS_RELIA_C_Q	$0.2174^{\downarrow}$	$0.4799^{\downarrow}$	$\mathbf{0.0558}^{\uparrow}$	$0.2528^{\downarrow}$	$0.1971^{\downarrow}$
PAS_RELIA_C_CS	$0.2427^{\downarrow}$	0.5143	$\mathbf{0.0645}^{\uparrow}$	$0.3116^{\downarrow}$	0.2472
PAS_RELIA_C_IS	$0.2176^{\downarrow}$	$0.4783^{\downarrow}$	$\mathbf{0.0636}^{\uparrow}$	$0.2711^{\downarrow}$	0.2075
Unsupervised methods					
PAS_RELIA_S_BN	$0.2700^{\downarrow}$	**0.5437**	$\mathbf{0.0614}^{\uparrow}$	0.3624	**0.3010**
PAS_RELIA_S_BS	$0.2722^{\downarrow}$	0.5391	$\mathbf{0.0767}^{\uparrow}$	0.3709	**0.2942**
PAS_RELIA_S_RN	$0.2584^{\downarrow}$	0.5428	$\mathbf{0.0533}^{\uparrow}$	0.3611	**0.3077**
PAS_RELIA_S_RS	$0.2657^{\downarrow}$	**0.5441**	$\mathbf{0.0571}^{\uparrow}$	**0.3990**	**0.3419**

Table 3.5: Passage reliability estimation results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

value for the ad-hoc retrieval task. All variants lead to global compatibility scores higher than that of the baseline, and provide statistically better compatibility harmful scores. In terms of CMAP, NDCG and compatibility helpful these methods are weaker. This suggest that, in general, these methods are better at downgrading harmful results but not so good at finding helpful results. The most robust method is `PAS_RELIA_S_RS`, which outperforms the baseline in nearly all metrics. This variant is based on a RoBERTa Large model obtained from the NLI and STSB datasets.

Note also that all variants show poor CMAP scores, while the NDCG scores (particularly those obtained with the unsupervised methods) are more competitive. CMAP (and Mean Average Precision, on which CMAP depends) is a measure influenced by how precision evolves over the entire ranking, while NDCG is a measure more oriented to high precision because it incorporates a discounting factor for relevant documents that grows with the position. NDCG has been recognized as a metric that reflects user behaviour well (e.g., web users rarely inspect a full ranking of results). In our case, NDCG is a more important measure, not only because

	Rprec Incorrect
Reference	
PAS_REL	0.1096
Supervised methods	
DOC_REL + PAS_RELIA_C_Q	0.0742$^{\downarrow}$
PAS_REL + PAS_RELIA_C_Q	0.0725$^{\downarrow}$
DOC_REL + PAS_RELIA_C_CS	0.0902
PAS_REL + PAS_RELIA_C_CS	0.0891
DOC_REL + PAS_RELIA_C_IS	0.0929
PAS_REL + PAS_RELIA_C_IS	0.0998
DOC_REL + PAS_REL + PAS_RELIA_C_Q	0.0771$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_CS	0.0990
DOC_REL + PAS_REL + PAS_RELIA_C_IS	0.1007
Unsupervised methods	
DOC_REL + PAS_RELIA_S_BN	0.1017
PAS_REL + PAS_RELIA_S_BN	0.1028
DOC_REL + PAS_RELIA_S_BS	**0.1155**
PAS_REL + PAS_RELIA_S_BS	0.1021
DOC_REL + PAS_RELIA_S_RN	0.1015
PAS_REL + PAS_RELIA_S_RN	0.0934
DOC_REL + PAS_RELIA_S_RS	0.1037
PAS_REL + PAS_RELIA_S_RS	0.0979
DOC_REL + PAS_REL + PAS_RELIA_S_BN	**0.1126**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1191**
DOC_REL + PAS_REL + PAS_RELIA_S_RN	0.1078
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.1171**

Table 3.6: CombSUM results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

it reflects a typical high-precision search scenario but also because it handles graded relevance (while CMAP/MAP only incorporate a binary notion of relevance).

3.6.3 Score Fusion

Having analyzed the individual effect of document relevance, passage relevance and passage reliability signals, we study now the effectiveness of combining multiple signals. Accordingly, we compare a selection of appropriate unsupervised rank fusion methods and learning-to-rank techniques.

Unsupervised Rank Fusion: CombSUM

Table 3.6 shows the results for the total recall task. Combining document/passage relevance with reliability estimation methods based on supervised techniques (second block of the table) leads to poor performance. The supervised strategy that trains with the incorrect sentence is again the best choice. However, its performance (`DOC_REL + PAS_REL + PAS_RELIA_C_IS` row) remains lower than that of the baseline. Combining document/passage relevance with reliability estimation methods based on unsupervised techniques (third block of the table) leads to more effective fusion variants. Several combinations outperform the baseline in terms of

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_RELIA_S_RS	0.2657	0.5441	0.0571	0.3990	0.3419
Supervised methods					
DOC_REL + PAS_RELIA_C_Q	0.2332$^{\downarrow}$	0.5033$^{\downarrow}$	0.0625	0.3088$^{\downarrow}$	0.2464$^{\downarrow}$
PAS_REL + PAS_RELIA_C_Q	0.2372$^{\downarrow}$	0.5009$^{\downarrow}$	0.0707	0.3086$^{\downarrow}$	0.2379$^{\downarrow}$
DOC_REL + PAS_RELIA_C_CS	0.2474	0.5149	0.0779	0.3589	0.2810$^{\downarrow}$
PAS_REL + PAS_RELIA_C_CS	0.2494$^{\downarrow}$	0.5126$^{\downarrow}$	0.0783	0.3544$^{\downarrow}$	0.2762$^{\downarrow}$
DOC_REL + PAS_RELIA_C_IS	0.2344$^{\downarrow}$	0.4958$^{\downarrow}$	0.0662	0.2944$^{\downarrow}$	0.2282$^{\downarrow}$
PAS_REL + PAS_RELIA_C_IS	0.2363$^{\downarrow}$	0.4983$^{\downarrow}$	0.0741	0.3226$^{\downarrow}$	0.2485$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_Q	0.2492	0.5158	0.0852	0.3431$^{\downarrow}$	0.2579$^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_C_CS	0.2583	0.5185	0.0853	0.3720	0.2868
DOC_REL + PAS_REL + PAS_RELIA_C_IS	0.2423$^{\downarrow}$	0.4999$^{\downarrow}$	0.0813	0.3265$^{\downarrow}$	0.2452$^{\downarrow}$
Unsupervised methods					
DOC_REL + PAS_RELIA_S_BN	**0.2821**$^{\uparrow}$	**0.5663**	0.0818	**0.4024**	0.3209
PAS_REL + PAS_RELIA_S_BN	**0.2754**$^{\uparrow}$	**0.5455**	0.0767	0.3844	0.3077
DOC_REL + PAS_RELIA_S_BS	**0.2693**	0.5317	0.1031$^{\downarrow}$	0.3850	0.2818
PAS_REL + PAS_RELIA_S_BS	**0.2837**$^{\uparrow}$	0.5423	0.0908$^{\downarrow}$	0.3974	0.3066
DOC_REL + PAS_RELIA_S_RN	**0.2774**$^{\uparrow}$	**0.5593**	0.0736	**0.4089**	0.3353
PAS_REL + PAS_RELIA_S_RN	**0.2728**	**0.5490**	0.0708	0.3896	0.3188
DOC_REL + PAS_RELIA_S_RS	**0.2774**$^{\uparrow}$	**0.5593**	0.0736	**0.4089**	0.3353
PAS_REL + PAS_RELIA_S_RS	**0.2753**$^{\uparrow}$	**0.5486**$^{\uparrow}$	0.0698$^{\downarrow}$	**0.4177**$^{\uparrow}$	**0.3479**
DOC_REL + PAS_REL + PAS_RELIA_S_BN	**0.2879**$^{\uparrow}$	**0.5648**$^{\uparrow}$	0.0954$^{\downarrow}$	**0.4162**	0.3209
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.2791**$^{\uparrow}$	0.5409	0.1104$^{\downarrow}$	**0.4039**	0.2935
DOC_REL + PAS_REL + PAS_RELIA_S_RN	**0.2859**$^{\uparrow}$	**0.5616**$^{\uparrow}$	0.0869$^{\downarrow}$	**0.4281**	0.3413
DOC_REL + PAS_REL + PAS_RELIA_S_RS	**0.2874**$^{\uparrow}$	**0.5571**	0.0829$^{\downarrow}$	**0.4370**$^{\uparrow}$	**0.3541**

Table 3.7: CombSUM results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

R-precision. However, no improvement is statistically significant.

For the ad-hoc retrieval task, the results show a similar trend. Combinations involving the supervised methods (second block, Table 3.7) show no benefit or even yield performance statistics that are statistically worse than those of the baseline. Fusion variants with unsupervised methods (third block, Table 3.7), instead, tend to produce improvements over the baseline (and many of them are statistically significant). Remarkably, the fusion of passage relevance with reliability estimation from the RoBERTa-Large-STSB model (`PAS_REL + PAS_RELIA_S_RS` row) and the fusion of document and passage relevance with reliability estimation from the RoBERTa-Large-STSB model (`DOC_REL + PAS_REL + PAS_RELIA_S_RS` row) show consistent improvements in terms of CMAP, NDCG and compatibility helpful. These methods are weaker in terms of compatibility harmful. The improvement of helpful-related metrics (CMAP, NDCG, compatibility helpful) usually comes at a cost of damaging harmful-related statistics (because we often move more on-topic documents to higher positions in the rankings and some of them might be harmful). This tradeoff between compatibility harmful and helpful is something we will discuss shortly and will be the subject of further analysis in Section 3.7. However, avoiding harm should not be our single criterion, as it would be trivial to achieve a system with perfect harmful scores (simply retrieving no documents would result in no harm produced).

	Rprec Incorrect
Reference	
PAS_REL	0.1096
CombSUM	
DOC_REL + PAS_RELIA_S_BS	**0.1155**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1191**
Borda Count	
DOC_REL + PAS_RELIA_S_BS	**0.1147**
DOC_REL + PAS_REL + PAS_RELIA_S_BS	**0.1222**
Leraning-to-Rank	
DOC_REL + PAS_RELIA_S_BS	$0.0621^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_S_BS	$0.0786^{\downarrow}$

Table 3.8: Borda Count and Learning-to-Rank results (only unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not.

	CMAP	NDCG	Comp. harmful	Comp. helpful	Compatibility
Reference					
PAS_RELIA_S_RS	0.2657	0.5441	0.0571	0.3990	0.3419
CombSUM					
PAS_REL + PAS_RELIA_S_RS	$\mathbf{0.2753}^{\uparrow}$	$\mathbf{0.5486}^{\uparrow}$	$0.0698^{\downarrow}$	$\mathbf{0.4177}^{\uparrow}$	**0.3479**
DOC_REL + PAS_REL + PAS_RELIA_S_RS	$\mathbf{0.2874}^{\uparrow}$	**0.5571**	$0.0829^{\downarrow}$	$\mathbf{0.4370}^{\uparrow}$	**0.3541**
Borda Count					
PAS_REL + PAS_RELIA_S_RS	$\mathbf{0.2894}^{\uparrow}$	$\mathbf{0.5529}^{\uparrow}$	$0.0892^{\downarrow}$	**0.4278**	0.3386
DOC_REL + PAS_REL + PAS_RELIA_S_RS	$\mathbf{0.2866}^{\uparrow}$	**0.5514**	$0.1133^{\downarrow}$	$\mathbf{0.4400}^{\uparrow}$	0.3267
Learning-to-Rank					
PAS_REL + PAS_RELIA_S_RS	$0.2022^{\downarrow}$	$0.4503^{\downarrow}$	$0.0312^{\uparrow}$	$0.1546^{\downarrow}$	$0.1234^{\downarrow}$
DOC_REL + PAS_REL + PAS_RELIA_S_RS	$0.2028^{\downarrow}$	$0.4494^{\downarrow}$	$0.0292^{\uparrow}$	$0.1505^{\downarrow}$	$0.1213^{\downarrow}$

Table 3.9: Borda and Learning-to-Rank results (only unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value.

Unsupervised Rank Fusion: Borda Count

The fusion experiments reported above were repeated for a second type of unsupervised fusion strategy: Borda Count. Overall, Borda count yielded similar results compared with the results obtained with CombSUM. For the sake of simplicity, we only report the effectiveness of Borda Count for the most effective fusion variants. In Tables 3.8 and 3.9, the reader can observe the relative merits of CombSUM (second block) against Borda Count (third block). This comparison does not reveal a clear winner. It seems that, for combining these pieces of evidence, the potential advantage of manipulating scores (CombSUM) does not translate into practical improvements in effectiveness.

Learning-to-rank

A second class of combination strategy consists of applying learning to rank methods. Given some training examples where we know the query-document scores[12] and the target variable (reliability of the document), we build a classifier that learns to combine the individual features. This strategy requires to further split the training queries into two subsets, where one subset is used to build the supervised models (if required) and the other subset is used by L2R to learn the combination of features. We set aside 5 queries for learning the combination. Observe that supervised methods (e.g., `PAS_RELIA_C_Q`): i) use the first subset of queries to learn the reliability estimation model, ii) the resulting reliability classifier predicts the reliability score for each document in the ranking of the second subset of queries, iii) the reliability scores together with the other query-document features are fed to the L2R model that learns the combination method, and iv) the learnt combination approach is run against the queries in the test fold. Unsupervised methods (e.g., `PAS_RELIA_S_BN`), instead, do not employ the first subset of queries: i) the unsupervised reliability estimation model predicts the reliability score for each document in the ranking of the second subset of queries, ii) the reliability scores together with the other query-document features are fed to the L2R model that learns the combination method, and iii) the learnt combination approach is run against the queries in the test fold.

We performed pointwise L2R, which predicts the value of the target variable for every single document. Results for both subtasks are shown in Tables 3.8 and 3.9 (last blocks). L2R does not give an added value over simpler fusion strategies. It seems that, with the available training queries, L2R methods are not able to learn a combination function that extrapolates to unseen queries. To further prove this point, Figure 3.8 represents the logistic regression weights obtained from each training fold (both subtasks are shown in the graph). It can be observed that the weight assigned to each signal varies enormously among folds. In total recall, the document relevance signal is always the feature assigned with lowest weight but its importance compared with the other two signals varies significantly over the three folds. In the ad-hoc retrieval task the situation is totally different: the passage relevance signal gets the lowest weights while document relevance and passage reliability show a erratic trend over the three folds. These plots support our hypothesis about the poor generalisation capability of the L2R algorithm. The graph clearly shows that the learned logistic regression models have high variance and we would need many more training examples to build a reliable combination model.

In search technologies, it is well known that there is a wide variability in the characteristics of queries. For example, some queries find a large number of relevant documents while other queries have few documents that are on-topic. This high variance makes that L2R methods would require a large corpus of training examples in order to build a robust combination ap-

[12] for a 3-feature combination we would have the document relevance score, passage relevance score and passage reliability score

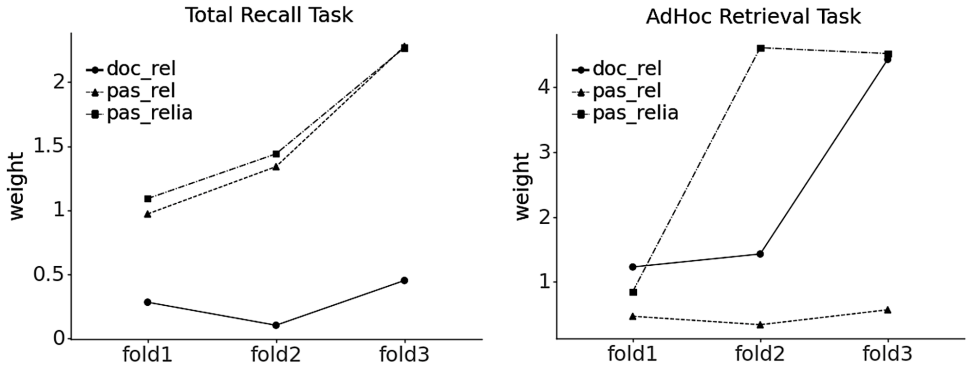


Figure 3.8: Logistic regression weights obtained from each training fold.

proach. But this luxury cannot be afforded in this misinformation detection task, where labelled data does not abound. In general web search scenarios, massive examples of topics and the associated clickthrough data are available to the search engine [32] and, thus, popular web retrieval engines can make good use of L2R strategies [104]. We focus instead on a more specific task where, in most of the cases, it is critical to avoid the spread of misinformation at early stages. This requires working with few training examples.

Comparison with external baselines

A way to put these results in context is to compare them with other studies using the same tasks and datasets. To that end, we consider here the participants in the TREC 2020 Health Misinformation competition[13]. For each task, we report the performance of the winner team (Best run block), the performance of the best run of the teams that ranked 2nd and 3rd (Other teams' runs block), the median performance of all runs (Median of all runs block) and the performance of some of our variants (last block).

For the total recall task, results are shown in Table 3.10. All our best performers (except one) are above the median *Rprec* of the submitted runs. Moreover, our top performer is better than 78% of the proposed solutions. The performance of the best run, KU from the University of Copenhagen [100], is higher than ours but their solution is based on a supervised model that was fed with external data, whereas our top performers are fully unsupervised.

For the ad-hoc Retrieval task (see Table 3.11), our improvement over the median increases, and our top performer is better than 88% of the submitted runs. Regarding the winner solution (H2oloo team, from the University of Waterloo), we obtain comparable compatibility Helpful results but we retrieve more harmful documents. This tradeoff will be further discussed in the next section. Nevertheless, it must be noticed again that the H2oloo solution is based on a supervised model trained with external data (and, in this case, obtained from a huge corpus, the

[13] To make results comparable with the other studies, our strategies had to be recomputed using all 50 topics assessed in the TREC task (instead of the 31 topics considered in Section 3.5). In this case, significance tests could not be performed because we have only these teams' mean scores (per-query results are not available).

	Rprec Incorrect
Best run	
KU (University of Copenhagen)	0.1300
Other teams' runs	
UWaterlooMDS (University of Waterloo)	0.1040
vohcolab (Universidade NOVA de Lisboa)	0.1030
Median of all runs	
MEDIAN RUN	0.0976
Our top performers	
DOC_REL + PAS_REL + PAS_RELIA_S_RS	0.1180
PAS_REL	0.1170
PAS_REL + PAS_RELIA_S_RS	0.0990
PAS_RELIA_S_RS	0.0801

Table 3.10: Comparison of official TREC 2020 runs and our best performers for the total recall task (data extracted from [42]).

	Comp. harmful	Comp. helpful	Compatibility
Best run			
H2oloo (University of Waterloo)	0.0160	0.4900	0.4740
Other teams' runs			
Webis (Bauhaus-Universität Weimar and Martin-Luther-Universität Halle-Wittenberg)	0.0520	0.3340	0.2820
KU (University of Copenhagen)	0.1210	0.4010	0.2800
Median of all runs			
MEDIAN RUN	0.0747	0.3337	0.259
Our top performers			
DOC_REL + PAS_REL + PAS_RELIA_S_RS	0.0825	0.4745	0.3920
PAS_REL + PAS_RELIA_S_RS	0.0711	0.4537	0.3826
PAS_RELIA_S_RS	0.0587	0.4209	0.3622
PAS_REL	0.1210	0.4370	0.3160

Table 3.11: Comparison of official TREC 2020 runs and our best performers for the ad-hoc retrieval task (data extracted from [42]). For the sake of simplicity we only report here the official metric by which the participating solutions were ranked (the difference between compatibility values).

T5-3b model [136]), while our top performers did not resort to supervision.

3.7 DISCUSSION

It is important to analyse the trade-off between the helpful and harmful compatibility results of the proposed solutions. Figure 3.9 plots some representative variants at the point where the X value corresponds with its compatibility helpful and the Y value corresponds with its compatibility harmful. Ideally, we want the system to be positioned at the bottom right of the graph because the main goal consists of minimising the retrieval of harmful results without damaging the retrieval of helpful documents.

We analyse here a representative set of variants, which includes a variant based only on document relevance (`DOC_REL`), a variant based only on passage relevance (`PAS_REL`), a variant based only on passage reliability (`PAS_RELIA_S_RS`) and four fusion variants (two of them based

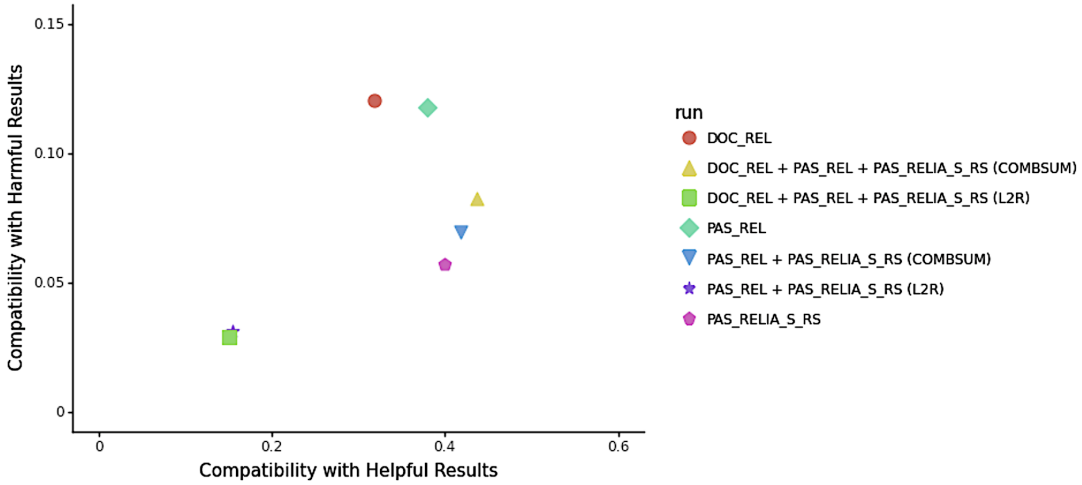


Figure 3.9: ad-hoc results: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a certain level of helpfulness, less harm is prefered.

on COMBSUM and two of them based on L2R). This analysis helps to further clarify some of our main findings.

The plot shows three main clusters of variants. The L2R variants, which yield low compatibility harmful and low compatibility helpful, the two relevance-based variants (`DOC_REL` and `PAS_REL`), which have high compatibility harmful and medium-to-high compatibility helpful, and the three remaining variants (`PAS_RELIA_S_RS` and the two COMBSUM variants), which have high compatibility helpful and medium compatibility harmful.

First, the two L2R alternatives, which are clustered at a low helpful and low harmful area, have limiting retrieval capabilities. Although they retrieve few harmful contents, their ability to bring helpful documents is clearly suboptimal. As argued above, we would need much more training data in order to make the most of these learning-based combinations.

Second, the initial retrieval of documents (`DOC_REL`) finds many harmful results. Given a document relevance ranking, we clearly need additional ingredients to decrease the retrieval of harmful documents and increase the retrieval of helpful contents. Passage-level relevance represents a first step in this direction. Compared with `DOC_REL`, `PAS_REL` improves the retrieval of helpful documents and, at the same time, slightly decreases the retrieval of harmful contents. This suggests that focusing on the most relevant extracts is beneficial to identify the most helpful webpages.

Third, `PAS_RELIA_S_RS`, `PAS_REL+PAS_RELIA_S_RS (COMBSUM)`, and `DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM)` are clearly the most solid choices, as they outperform the two relevance-based variants in both compatibility measures (lower compatibility harmful and higher compatibility helpful compared to `DOC_REL` or `PAS_REL`). If we want to fare on the conservative side, we could choose `PAS_RELIA_S_RS`: it retrieves fewer helpful documents but it also results in less damage. On the other hand, if we want higher recall of helpful documents then `DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM)` would be our preferred choice: it re-

trieves more helpful webpages at the cost of presenting more harmful contents in the rankings. In practice, the selection of one of these methods would depend on the specific user task and his/her willingness to weight on helpfulness or harmfulness. For example, a website moderator willing to thoroughly inspect the presence of helpful and harmful contents within his/her site would probably prefer `DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM)`. But if the goal is to identify the most reputed contents about a given topic (e.g., to label them as useful suggestions) then `PAS_RELIA_S_RS` would be a more cost-effective strategy (similar helpful results compared with `DOC_REL+PAS_REL+PAS_RELIA_S_RS (COMBSUM)` and `PAS_REL+PAS_RELIA_S_RS (COMBSUM)`, but lower harmful results).

In general, we found the following tendency: the better our retrieval systems are in terms of compatibility helpful, the more harmful documents are also found. And the other way around, if we decrease the retrieval of harmful webpages it is often at the cost of decreasing the retrieval of helpful webpages. It is quite difficult to find an artifact that substantially improves both dimensions. In any case, our goal in the future is to continue studying the specifics of this challenging task and conduct research on new features or strategies oriented to show a good balance between helpfulness and harmfulness. We are also interested in designing novel thresholding strategies adapted to this retrieval problem (e.g., given a ranked set of webpages determine the ideal cutoff position taking into account both dimensions).

The main takeaways could be summarised as follows:

- Focusing on the most relevant passages of documents leads to benefits that are modest but promising. This passage-relevance approach tends to improve helpfulness and decrease harmfulness.

- Estimating passage reliability also helps. However, we found that there is a substantial difference between opting for a supervised or an unsupervised estimation, being the latter the best performer. At the early stages of an information outbreak (and COVID-19 is a clear case), the availability of topically-related training data is scarce and our results clearly demonstrate that, under this stringent scenario, unsupervised reliability estimation seems to be a good choice. However, there are also semi-supervised learning or transfer learning techniques, which could be considered to further support this task. In the near future, we plan to explore the role of semi-supervised models or transfer learning models for these tasks.

- Simple score fusion techniques like CombSUM have been demonstrated to outperform L2R strategies for this task, which would require more training data to reach their full potential.

3.8 Participation in the TREC 2021 and TREC 2022 Health Misinformation (HM) Tracks

We also tested this technology in the context of the TREC Health Misinformation Tracks in 2021 and 2022. In contrast with the 2020 edition, these years the track focused on general health information topics.

In the 2021 edition, we obtained our best global results ending in a meritorious third position. Our top performing solution was based on the retrieval system presented in this chapter and it consisted of an initial document level BM25 search plus a MonoT5 passage re-ranking of the top 100 retrieved documents using the correct sentence derived from the question and answer fields [58].

In the 2022 edition, we continued taking part in the adhoc retrieval task, and we also joined a new answer prediction challenge. This task aimed at predicting the correct response to a medical query. For this task, our proposal was based on distilling knowledge from search engines result pages (SERPs) and, additionally, exploiting the language understanding capabilities of GPT-3-based models [57]. This was the seed for further research into the reliability of the new Large Language Models (LLMs) in providing reliable medical advice. This will be discussed in Chapter 4.

3.9 Final remarks

In this chapter, we have presented a thorough study on the signals and combination methods that are potentially helpful in the task of identifying health-related misinformation. We contributed with:

- A complete multistage retrieval system whose goal is to discern between reliable and unreliable contents. This technological solution is available to be reused or adapted[8] by researchers interested in misinformation, web moderators, vertical search engine creators, or other potential stakeholders.

- A comparative study that empirically validated the potential of the platform for a socially worrying case, COVID-19 misinformation. Our analysis has assessed the effect of search-based stages, at document and passage level, reliability estimators based on supervised and non-supervised models and different fusion strategies. Every stage that we included in our system improved the overall performance, and in some cases, significantly. The fusion or combination of multiple forms of evidence (document relevance, passage relevance and passage reliability) led to the most efficient misinformation estimation methods.

– The top-performing variants have competitive performance when compared with state-of-the-art methods (and, particularly, with respect to the solutions submitted to the TREC 2020 Health Misinformation Track).

– The trade-off between retrieval of helpful and harmful contents has been analysed in depth. The results reflect that certain signals help more in finding more helpful documents, while others are more prone to limiting the retrieval of harmful contents. However, it is still challenging to find a combination that improves both aspects. The choice of one instance of the system over another would depend on the specifics of the search task.

The findings of this study have to be seen in light of some limitations. The primary limitation to the generalisation of these results is the test collection. Although we performed experiments with two different search tasks, the document collection was the same and the number of available search topics is limited. In the near future we want to extend the empirical validation to new datasets and larger sets of topics. The second limitation concerns the signals analysed. This study has been confined to text-based search or classification signals. It would be interesting to test other types of features, such as those based on network signals (e.g., link-based reputation of the web sources) or interaction/social signals (e.g., effect of the publications on Internet users).

As future work, we also want to further understand the trade-off between harmful and helpful compatibility, and design strategies to determine the ideal cutoff of a ranked list adapted to this problem [6]. We are currently working on additional NLP techniques to be included in our pipeline. For example, we are working with other unsupervised techniques for further removal of noisy contents [171] (see Section 5.2 of Chapter 5 of this manuscript). In this respect, we will carefully consider recent advances in clustering [53] and how to effectively employ clustering algorithms to further improve misinformation detection. For example, it will be interesting to exploit clustering techniques for organizing the retrieved results into groups of helpful and harmful pages.

Other possible lines of future work include the application of rule-based techniques and new feature selection algorithms [153] to filter rumours or misinformation in the health domain. Related to this, we want to further analyse recent affective computing and sentiment analysis models [111, 82] and study how to employ them to define new features or signals for the task of misinformation detection. Finally, it is also important to work towards the explainability of the proposed solutions. Some parts of our multistage system are based on deep learning models that are black-box techniques and, thus, hard to interpret. We want to learn from recent advances in this area [27] and see how to adapt these proposals to our application domain.

4 Reliability of LLMs in Providing Medical Advice

The phenomenal development of LLMs in the past two years has had a global impact on Information Technologies. The traditional use of search engines to retrieve relevant documents is rapidly being displaced by conversational AI services that, instead of providing web results, directly generate answers to users' information needs. This motivated us to explore the LLMs' abilities in addressing health-related information needs. This is a natural complement of the research presented in the previous chapters, which focused on the most traditional document retrieval information access paradigm.

The emergence of Large Language Models (LLMs) has induced significant improvements in performance on various Natural Language Processing (NLP) tasks such as Language Understanding [139], Text Generation [140], and Machine Translation [168]. The appearance of BERT [50], GPT-2 [141], and GPT-3 [25], among others, has accelerated the development of LLMs. In November 2022, the release of OpenAI's Chat-GPT[1] marked a significant milestone in the evolution of LLMs. This new model pushed the boundaries of what was thought feasible in terms of producing coherent and human-like text [63]. ChatGPT's growth has been nothing short of phenomenal, as it is estimated to have reached 100 million monthly active users in January 2023, becoming the fastest-growing application in history. This disruptive event shifted the focus of AI research to LLMs, sparking renewed interest in learning more about their capabilities and knowledge reasoning. These developments have motivated extensive research into the potential of LLMs and their impact on various NLP applications [98, 85].

With the increasing reliance of users on online medical information [65], the reliability of these models to provide accurate medical advice needs to be put under scrutiny. The potential consequences of incorrect health-related information can result in personal harm [162, 131]. Hence, there is an urgent need to investigate the capabilities and limitations of LLMs in providing medical advice. The evaluation of the robustness of these models in this critical domain is of utmost importance. In addition, it must be taken into account that the performance of LLMs is highly dependent on the prompt and context provided by the questioner [25, 85, 103].

In this chapter, we present a study aimed at estimating the reliability of LLMs in providing medical advice, exploring their potential to support end-users who submit health-related infor-

[1] https://chat.openai.com/chat

mation requests. To that end, we compare multiple LLMs (including recently released models, such as GPT-4) and examine their performance on a range of medical questions. Our evaluation considers a wide range of context and prompt conditions and we further discuss the potential challenges and implications of using LLMs for medical information needs.

The contents of this chapter are extracted from the following publication:

> Fernández-Pichel, M.[a], Losada, D.E. *a*, Pichel, J.C. *a*. (2023). *Large Language Models for Binary Health-Related Question Answering: A Zero- and Few-Shot Evaluation*. Submitted to the 46th European Conference on Information Retrieval (ECIR '24).
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

4.1 Background

Current LLMs have great potential for providing medical information. However, their reliability for such critical information service remains largely unknown. Previous studies have explored various techniques to estimate the general knowledge of LLMs and how to prompt them to generate more reliable responses. For instance, Petroni et al. [130] proposed the LAMA benchmark, a set of language understanding tasks designed to evaluate the factual and common sense knowledge of LLMs. They demonstrated that LLMs retain factual knowledge without any fine-tuning, showing their potential for question answering tasks.

Brown and colleagues [25] introduced GPT-3 and demonstrated its capacity for a variety of tasks under a few-shot learning setting. The paper deeply discusses the use of natural language prompts, highlighting the role of effective prompting in distilling the model's knowledge. Other studies [85] tried to optimise knowledge discovery in LLMs by generating high quality prompts (manual or automatic) and by exploiting ensemble methods.

Liu et al. [102] focused their efforts on another critical aspect, the optimal configuration of in-context examples to enhance GPT-3's few shot capabilities. They found that this is specially crucial in Text Generation tasks. Another recent study [98] performed a holistic evaluation of different models, prompts, metrics and tasks. In early 2023, Liu and colleagues [103] published a systematic review about different prompting methods and the appearance of ChatGPT has stimulated targeted studies to gauge the model's knowledge and utility for a number of tasks [22, 158, 4].

A recent study [179] analysed the impact of prompts in health information seeking. However, the study was confined to a single LLM (ChatGPT) and the main goal was to evaluate prompts that incorporate supporting and contrary evidence obtained from a search engine.

In this chapter, we contribute with a systematic evaluation on the reliability of current generative models in providing medical answers. To this aim, we prompted the models with standard

medical questions and several manually designed contexts, and we empirically evaluated the quality of the answers (compared to current medical practice). We also report our endeavours to comprehend the models' predictions (e.g., about a medical treatment) and tried to understand the possible limitations and risks these technologies may have.

4.2 Experimental design

Through the conducted experiments, we try to answer the following research questions:

- To what extent do LLMs provide correct medical advice?
- How different LLMs (GPT3-based models, ChatGPT, GPT4, and Flan T5) compare in terms of effective response to medical questions?
- To what extent does the provided context make that these models provide the right answer?
- Do these models improve when presented with a few in-context examples?
- What type of errors do these LLMs make when prompted with medical questions?

4.2.1 Models

For a rigorous experimentation, we considered LLMs of different philosophies (e.g., open source vs proprietary), architectures, complexity and training data:

- **GPT-3, text-davinci-002** (d-002). It is a decoder-only architecture consisting of 175 billion parameters. It was trained on a wide range of web-crawled corpus, including the entire Wikipedia (with data up to June 2021). It has been tested in zero-shot and few-shot settings for several downstream NLP tasks where it showed impressive performance [25].
- **GPT-3, text-davinci-003** (d-003). It is an improved version of the previous model [25] that was built on top of InstructGPT [125]. InstructGPT models are fine-tuned with human feedback using reinforcement learning. Again, it was trained with data up to June 2021.
- **ChatGPT, gpt-3.5-turbo**. It is similar to InstructGPT, but it meant a paradigm shift towards more conversational interaction [62]. Its training data goes up to September 2021.
- **GPT-4, gpt-4-8k**. This particular bot was also designed for conversational purposes. It serves as a cutting-edge advancement in this field and surpasses ChatGPT's performance in various tasks that require human-like intelligence, such as passing an exam [123]. Its training data also goes up to September 2021.
- **Flan T5, flan-t5-xl** (FT5). This is a sequence-to-sequence model developed by Google. The model was fine-tuned by its creators using multiple datasets that consist of instructions, as documented in [40]. The datasets were procured from a open source repository, known as "Flan 2022" [107], which contains comprehensive information gathered up until the year 2022.

The first four models were tested through OpenAI's official Python API[2], while Flan T5 was tested through its Hugging Face implementation[3]. We set the models' temperature to 0, with the intention of minimising randomness in their responses. These experiments were run between February and April 2023.

To facilitate reproducibility, we provide a repository containing the code used and the results obtained for the different models[4].

4.2.2 Health-related questions

To assess the LLMs, we used the data from the TREC Health Misinformation (HM) Tracks of 2020, 2021 and 2022. As argued above, these collections consist of health-related queries, in the form of questions (e.g., can wearing masks prevent COVID-19?), and web documents. We utilised the questions and their ground truth answers (yes/no), which represent the best understanding of current medical practice. The 2020 questions are all related to COVID-19, while the 2021 and 2022 questions are general health information needs. The 2020 questions were released in mid 2020 and, thus, we cannot discard that the LLMs have seen this benchmark within their training data. The 2021 questions were released in mid July and, thus, these topics might have been available for ChatGPT and GPT-4 (but not for GPT-3, whose training ended earlier). The 2022 questions, instead, were made available after the construction of any of the LLMs. This therefore conforms an assorted set of health questions, with varying levels of difficulty for the models (depending on their exposure to this type of data and the level of specificity of the information needs).

4.2.3 Prompts

As we pointed out above, several studies have demonstrated that effective prompts have a direct effect on the quality of LLMs' output [85, 103]. In this study, we compared the performance of different prompts or conditions for the same medical questions:

- **no-context**: a prompt composed only of the medical question, i.e. "*Can Vitamin D cure COVID-19?*".
- **no-context**^{m} : The text "*The answer must be Yes or No*" concatenated after the question. The goal of the *prompt*^{m} variants is to force the LLM to give a yes/no answer. We expect these variants to be less conservative and, thus, more error-prone.

[2] https://openai.com/blog/openai-api
[3] https://huggingface.co/google/flan-t5-xl
[4] https://github.com/MarcosFP97/llm-health-advice-evaluation

	TREC HM 2020					TREC HM 2021					TREC HM 2022				
prompt	d-002	d-003	Chat-GPT	GPT4	FT5	d-002	d-003	Chat-GPT	GPT4	FT5	d-002	d-003	Chat-GPT	GPT4	FT5
no-context	0.84	0.9	0.82	0.79	0.24	**0.72**	0.74	0.68	0.66	0.44	0.76	0.76	0.76	0.86	0.56
no-context^m	0.69	0.91	0.84	**0.92**	**0.84**	0.54	0.74	0.64	**0.7**	**0.7**	0.56	0.72	0.86	0.78	0.56
non-expert	0.78	0.9	0.8	0.9	0.31	0.4	0.62	0.54	0.66	0.54	0.48	0.72	0.8	0.86	0.54
non-expert^m	**0.86**	**0.92**	0.8	0.74	**0.84**	0.68	0.58	0.58	0.6	0.6	0.74	**0.8**	0.78	0.8	0.56
expert	**0.86**	0.89	0.84	0.84	0.79	0.36	0.8	**0.7**	0.68	0.64	0.68	0.74	**0.9**	0.88	**0.74**
expert^m	0.84	0.9	**0.9**	0.88	**0.84**	0.7	**0.82**	0.68	**0.7**	**0.7**	**0.82**	**0.8**	0.88	**0.9**	0.6
avg	0.81	0.90	0.83	0.84	0.57	0.55	0.71	0.63	0.67	0.60	0.66	0.76	0.83	0.85	0.59
std	0.07	0.01	0.04	0.07	0.29	0.16	0.09	0.07	0.04	0.10	0.13	0.04	0.06	0.05	0.07

Table 4.1: Zero-shot experiments, proportion of correct answers of each model-prompt combination for the three TREC datasets. The best performing prompt for each model is marked in bold. The last two rows report the average performance and the standard deviation of each model across all prompts.

- **non-expert**: The text "*I am a non-expert user searching for medical advice online*" plus the corresponding question. This prompt might be representative of a regular user searching for medical advice.
- **non-expert^m**: the same prompt as before plus "*The answer must be Yes or No*".
- **expert**: The text "*We are a committee of leading scientific experts and medical doctors reviewing the latest and highest quality of research from PubMED. For each question, we have chosen an answer, either 'yes' or 'no', based on our best understanding of current medical practice and literature.*" plus the corresponding medical question. This prompt was designed by Waterloo's team in their participation in the TREC 2022 Health Misinformation track[5]. The rationale is to bias the LLM towards reputed contents associated to high quality sources.
- **expert^m**: "*The answer must be Yes or No*" added after the expert prompt.

It is also worth noting that conversational-oriented models, like ChatGPT and GPT-4, offer the possibility of prompting the *system role* and/or prompting the *user role*. For the sake of consistency with the rest of the models, we submitted the previous prompts to the user role and we used the default system role ("*You are a chatbot*").

Given the LLMs' response to the prompt, if the output of the model was affirmative or negative then it was automatically recorded as the model's answer. We did manual inspection for those cases where the textual response was somehow equivocal. We report the proportion of correct answers. Note that the rest of cases can be either incorrect answers or unanswered questions.

4.3 Results

4.3.1 Zero-shot experiments

As can be seen in Table 4.1, text-davinci-003, ChatGPT and GPT-4 are the best performing models. This is not surprising, as these three LLMs are the most modern and advanced models.

[5] https://trec-health-misinfo.github.io/

There are also some differences in performance among the prompts evaluated. The most robust context seems to be expert^m. We hypothesise that this is due to the inclusion of keyphrases such as "*research from PubMed*" or "*medical practice and literature*", which bias the model towards reputable sources of knowledge.

The effect of forcing the model to give a yes/no answer is clearly dependent on the type of prompt. In some cases, demanding a yes/no answer seems to be beneficial (e.g., non-expert^m vs non-expert and, to some extent, expert^m vs expert) but some other comparisons (e.g., no-context^m vs no-context) show a mixed-bag of results. The variance on the quality of responses is clearly an issue for some of the models. FT5 is clearly dependent on the type of prompt. Although the other models are relatively stable, they still exhibit notable variations depending on the input provided by the user. This is concerning, as a model's effectiveness can range from 90% of correct answers to $\approx$ 75% of correct responses. The overall levels of effectiveness are remarkable but, still, these inconsistencies are a cause of discomfort. Even adopting the most consistent prompt (expert^m) we observe concerning outcomes. For example, ChatGPT suffers from poor performance (68%) in the 2021 dataset.

The three datasets vary in their level of difficulty. The 2020 health questions are more specific (related to COVID-19) and it appears that this simplifies matters for the LLMs. A plausible explanation for this phenomenon could be that the models might have already been exposed to these health questions during their massive training, as stated in Section 4.2.2. Another explanation could be that the highly relevant and significant nature of COVID-19 as a topic might have motivated a specialised curation process for the relevant data. But this is just an speculation, as the creators of these LLMs hardly disclose specific details about the training corpora.

As part of our analysis, we employed McNemar's test [94] to assess the significance of the differences between the top-performing models. Between ChatGPT and GPT-4, we found no significant difference in 16 out of 18 comparisons (3 collections $\times$ 6 prompts). The pairwise comparisons d-003 vs ChatGPT and d-003 vs GPT-4 revealed more cases of statistical significance but, still, more than half of the compared instances yielded a no significance result. On the contrary, the weaker models, d-002 and FT5, differed significantly from the stronger models and, usually, the test marked the differences as statistically significant.

In Section 4.4, we further discuss the type of errors made by the models and make additional qualitative analyses.

4.3.2 Few-shot experiments

Let us go one step further and assess the effect of including in-context examples in the final performance. This empirical analysis was done with the test questions from TREC HM 2022, which were prompted with one-to-three demonstrations extracted from TREC HM 2021. We

model	prompt	0-shot	1 in-context example	2 in-context examples	3 in-context examples
d-002	no-context	**0.76**	0.58	0.58	0.6
	no-context^m	0.56	0.74*	0.7	**0.8***
	non-expert	0.48	0.42*	0.46	**0.5**
	non-expert^m	0.74	0.7	0.76	**0.78***
	expert	**0.68**	0.62*	0.64	**0.68**
	expert^m	0.82	0.72	0.82	**0.84**
d-003	no-context	0.76	0.7*	0.8	**0.82**
	no-context^m	**0.72**	0.7*	**0.72***	0.68*
	non-expert	0.72	0.68	0.68*	**0.8**
	non-expert^m	**0.8**	0.64*	0.66*	0.62*
	expert	0.74	0.62*	**0.78**	**0.78**
	expert^m	**0.8**	0.66*	0.72	0.66*
ChatGPT	no-context	0.76	**0.84**	0.8	0.82
	no-context^m	**0.86**	0.8	0.8	**0.86**
	non-expert	**0.8**	0.78	0.78	0.78
	non-expert^m	0.78	**0.86***	0.84	0.8
	expert	**0.9**	0.84	0.82	0.82
	expert^m	**0.88**	0.84	0.86	0.8
GPT-4	no-context	0.86	**0.92**	0.88	0.9
	no-context^m	0.78	**0.9**	0.88	0.88
	non-expert	**0.86**	0.8	**0.86**	0.84
	non-expert^m	0.8	0.86	0.86	**0.88**
	expert	0.88	**0.9**	**0.9**	0.86
	expert^m	**0.9**	**0.9**	0.88	0.88
FT5	no-context	0.56	0.62*	0.62*	**0.64***
	no-context^m	0.56	0.56*	**0.58**	**0.58**
	non-expert	0.54	**0.62***	0.6*	0.58*
	non-expert^m	0.56	0.56*	**0.62**	**0.62***
	expert	**0.74**	0.66	0.68*	0.68*
	expert^m	**0.6**	0.58	**0.6**	0.58

Table 4.2: Few-shot experiments, proportion of correct answers of each model-prompt combination with varying number of in-context examples. For each row, the best score is marked in bold and the symbol * marks those cases where McNemar's test ($\alpha = .05$) finds a difference between the corresponding model and its 0-shot equivalent.

randomly chose three pairs of *(medical question, correct answer*[6]*)* from the 2021 dataset as in-context examples and explored the effect of including from one to three before each of the prompts[7]. Note that the strict limitations on input size in present LLMs hinder the inclusion of a large number of in-context examples. Furthermore, past research [98] has shown that including an excessive number of in-context examples is ineffective.

As shown in Table 4.2, the effect of the in-context examples strongly depends on the model and type of prompt. The most effective models, such as ChatGPT or GPT-4, do not seem to require these demonstrations. And, regardless of the model, the most solid prompt (expert^m) does not get any added value from the inclusion of the examples. In terms of statistical significance (McNemar's test between the 0-shot case and each few-shot case), only the weakest models (e.g., FT5 or d-002) show some instances, mostly associated to simple prompts, where the inclusion of in-context examples was clearly beneficial.

[6] Yes or No

[7] To ensure comparability of results, we utilised the same sampled examples for all the variants under comparison.

4.4 Error Analysis

To further understand the errors made by the LLMs, we manually inspected the queries where none of the models provided a correct answer. This analysis was done for the best performing prompt (*expert*^{m}) and for the *no-context* prompt. We identified three main classes of errors:

- **Lack of knowledge about current medical consensus**. For instance, to the question "*Can Hydroxychloroquine worsen COVID-19?*", d-003 answered "*no, there is no evidence that hydroxychloroquine worsens covid-19...*" while medical evidence says otherwise[8]. A similar error is made for the question "*Can omega-3 treat borderline personality in women?*" (omega-3 has been found to be helpful but LLMs argue otherwise).

- **Lack of common sense knowledge**. Here, LLMs err due to a flawed interpretation of the question. For example, "*Can bleach prevent COVID-19? No, bleach should not be ingested...*". But the ground truth has the most obvious interpretation of this question (the use of bleach for surface disinfection can actually prevent COVID-19). A human would hardly interpret the question in the way that the LLM does.

- **Indecisive responses**. Here, LLMs tend to explain pros and cons (e.g., "*Are sit ups bad for you?*") and do not give a blunt response. These are not counted as correct responses but the LLM's output is arguably useful ("*Sit-ups can be both beneficial and harmful, depending on your individual circumstances and the way you perform the exercise...*").

The first class of errors is clearly the most problematic. For the second class (incorrect interpretation) the inquirer would probably rephrase the question to clarify the intended meaning and in the third class of cases the LLMs provide an output that is at least partially useful.

Note also that regular end-users are not likely to provide a solid and detailed prompt (e.g., to guide the LLM towards reputed medical contents). We need new tools that wrap the user's questions into proper contextual prompts. To that end, we have developed a preliminary version of a Python wrapper that encapsulates different types of health-related prompts and combines them with simple user requests. This tool also facilitates the reproducibility of our experiments[9].

4.5 Final remarks

In this chapter, we have conducted a thorough analysis of the use of LLMs for obtaining medical advice. We have demonstrated that the selected prompt strongly influences performance and, in some cases, including in-context examples can be helpful.

[8] FDA cautions against use of hydroxychloroquine or chloroquine for COVID-19.
[9] `https://github.com/citiususc/Smarty-GPT`

Although the effectiveness of these models is remarkable, there are still some concerning mistakes. For example, even the best models with the most solid prompts produce some answers that go against current medical knowledge. And, sometimes, LLMs show a lack of common sense knowledge and make incorrect interpretations of the question.

5 Tools for massive processing of online content

In the previous chapters we have described different scientific alternatives for the automatic detection of misinformation, including the application of classical document-level approaches, the design of a new multistage system that combines multiple passage-level and document-level forms of evidence, and the exploitation of new LLMs. However, the deployment of these predictive models at web scale requires the construction of technological solutions capable of monitoring and efficiently preprocessing large amounts of data. This chapter is dedicated to these challenges. It outlines how, during this doctoral research, we have developed demonstrators (oriented to specific use cases) that have given support to our studies and, additionally, innovative language analysis tools for cleansing textual content.

In Section 5.1, we introduce Social Minder, a modular platform for credibility analysis in Social Media. The main goal of this technological demonstrator was to put document-level estimators of credibility in practice. We explain the architecture of the platform and present a use case oriented to misinformation detection of contents related to COVID-19.

The analysis of large amounts of social media and web data not only presents challenges related to Big Data processing, but also introduces difficulties associated with the noise inherent in the contents. When handling web pages, we realised that the extraction of the core textual excerpts requires to remove many noisy and meaningless sentences. Thus, we developed a perplexity-based estimator that, using a reference language model, selects the sentences which should be removed from a given text. In Section 5.2, we explain this technique and demonstrate its utility for several downstream tasks (text classification, adhoc retrieval, and tag cleaning).

The contents of this chapter are extracted from the following publications:

> Fernández-Pichel, M.[a], Losada, D. E.[a], and Pichel, J. C[a]. (2022). *Social Minder: a Tool for Social Media Monitoring and its Use for Detecting COVID-19 Misinformation*. In 2nd Joint Conference of the Information Retrieval Communities in Europe, CIRCLE 2022. CEUR Workshop Proceedings. The publication is available at: `https://ceur-ws.org/Vol-3178/CIRCLE_2022_paper_01.pdf`.
>
> [a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

Fernández-Pichel, M.[a], Prada-Corral, M.[a], Losada, D.E.[a], Pichel, J. C.[a], and Gamallo, P.[a]. (2023). *An unsupervised perplexity-based method for boilerplate removal*. Natural Language Engineering, 1–18. The publication is available at: `https://doi.org/10.1017/S1351324923000049`

[a]Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Spain.

5.1 Social Minder

5.1.1 Background

Social Media analytical tools are often constrained to work from data provided by official APIs, as Batrinca and Treleaven showed in their thorough survey [12]. One of the advantages of Social Minder is that it allows massive extraction of tweets with its own crawler [112] and works with a modular and scalable architecture that can efficiently ingest large amounts of textual data (see Section 5.1.2).

Existing tools for social media monitoring, such as Social Mention, provide a rigid set of functionalities (e.g., general statistics about queries). Social Minder differs from these because it includes a real-time credibility estimation module with self-developed technology. This module was built following the lessons learned in our research and, particularly, from the experimental results detailed in Chapter 2 (Section 2.1).

Although there are some existing initiatives for real-time credibility analysis on Twitter [75], to the best of our knowledge, our platform is the first to integrate this functionality into a complete monitoring system expandable to other web sources, not only Twitter.

Related to our use case, the study by Sharma and colleagues [152] also addressed COVID-19 misinformation on Twitter. However, the main difference here lies in the way that misinformation is detected. These authors proposed a manual annotation technique, based on certain expressions and hashtags, while Social Minder incorporates an automatic algorithm, as decribed in Section 5.1.2.

5.1.2 Architecture

Social Minder was built on the top of eXtream [55], a Big Data framework that permits advanced users to design their own data processing topologies. Social Minder is an evolution oriented to the end-user, providing a dashboard for non-expert users. Its system architecture consists of a fixed consumption topology that interconnects several containerised modules (see Figure 5.1). The functionality of each module is briefly explained below:

– A Twitter crawler [112] that injects text streams into the topology. For a given query, it first tries to recover all historical tweets, and then starts to consume in real-time.

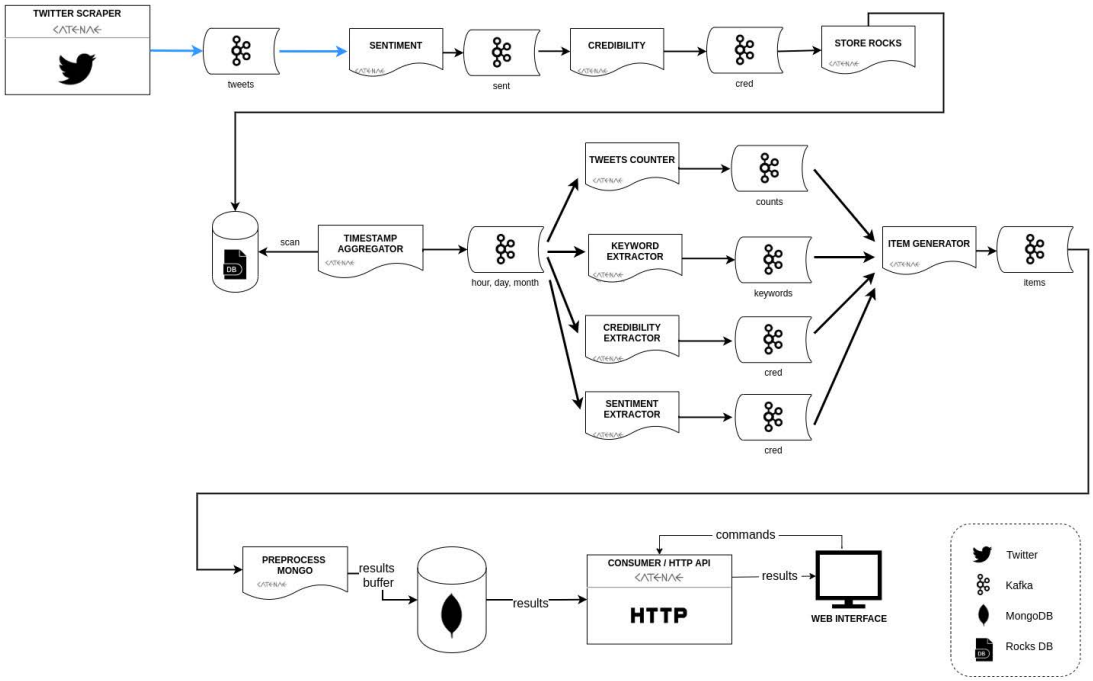


Figure 5.1: Social Minder architecture.

– A sentiment analysis module based on VADER [83], which incorporates rule-based classification technology.

– A credibility estimation module that uses a self-developed classification technology, based on our experimental results from Section 2.1. This is a classifier similar to the one we employed in our participation in the TREC HM 2020. It consists of a support vector machine trained on three credibility classification datasets [155, 86, 150]. Since the training data comes from the Web Search domain, only the web pages linked in the tweets are assessed for credibility. Tweets that do not contain any link are ignored.

– A timestamp-aggregator module that groups texts by different temporal granularities (*hour*, *day*, *month*) to perform the analysis.

– Four parallel computation modules that perform different statistical analysis tasks (count texts, extract keywords using *TF-IDF* techniques, compute aggregated sentiment and credibility) for all temporal granularities available.

– Two final modules that aggregate results and write them on permanent storage (Mongo DB database).

– A dashboard for non-expert end-users that shows statistics and graphs per query for different granularities.

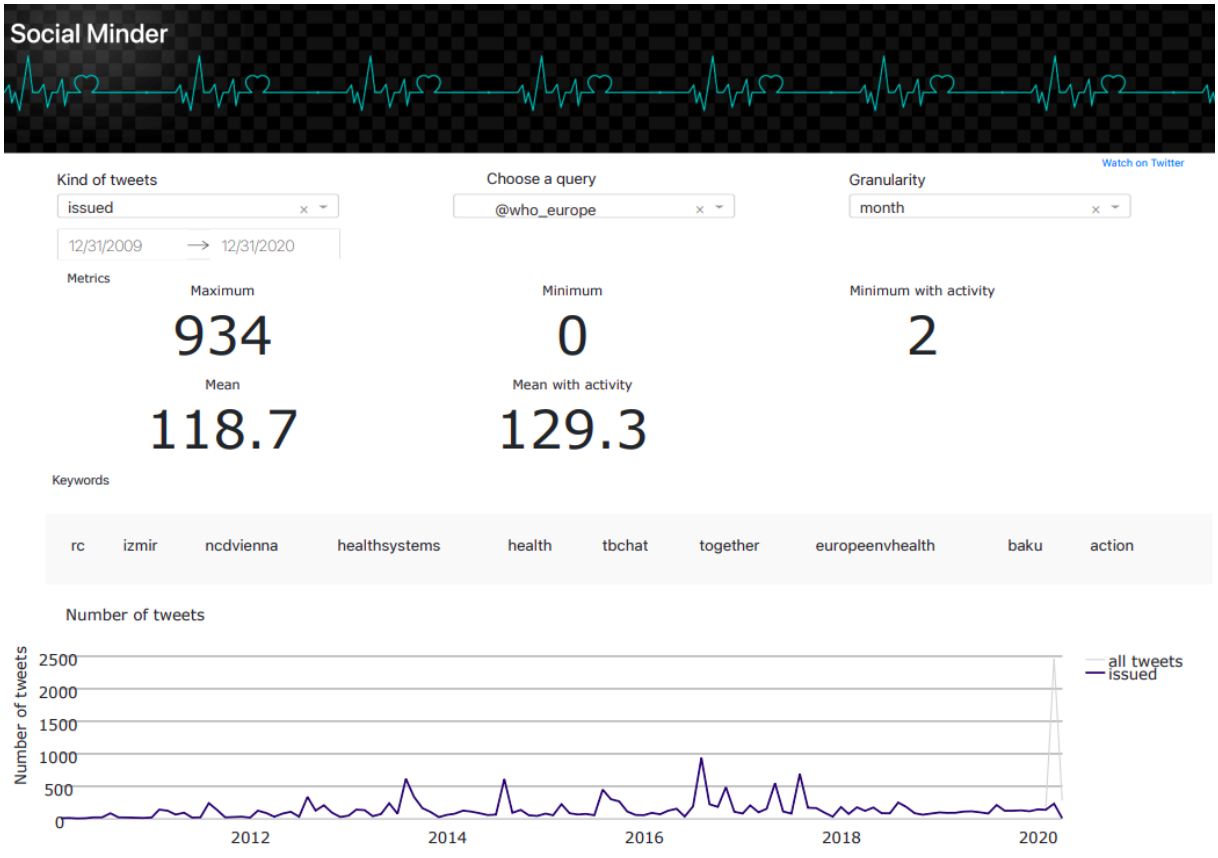


Figure 5.2: Social Minder dashboard (upper part).

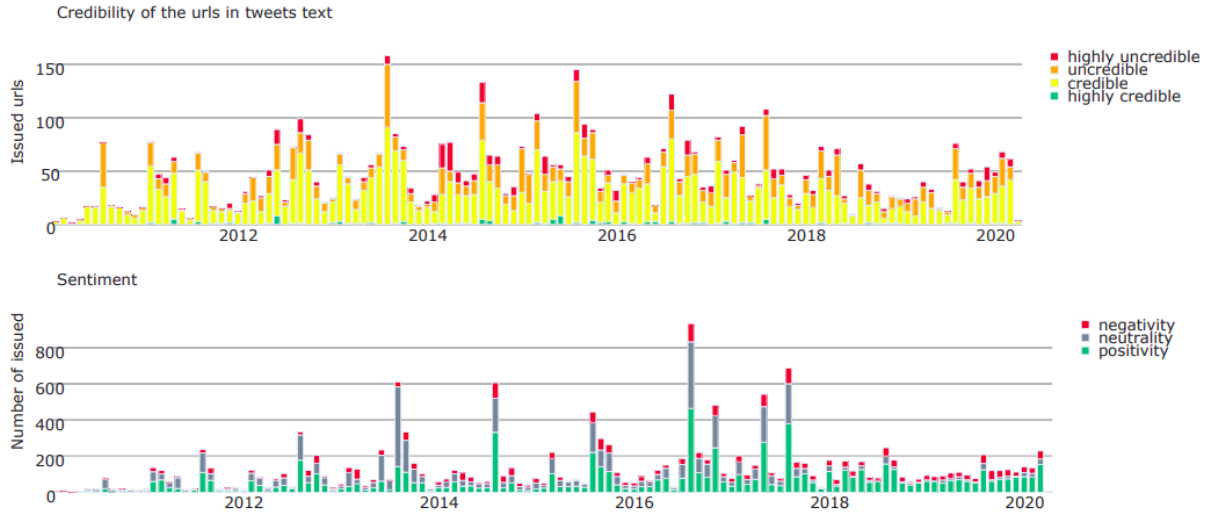


Figure 5.3: Social Minder dashboard (bottom part).

5.1.3 COVID-19 misinformation use case

Social Minder can serve multiple research or commercial purposes. For example, one can develop new social media applications by modifying the profiles of interest. This demo focuses on a use case of Social Minder oriented to monitor misinformation posted on Twitter about

COVID-19. We exemplify the tool with a dashboard associated to a sample of covid-related tweets obtained in 2020[1].

Social Minder allows to filter the Twitter stream either by account or by a textual filter. We illustrate this by considering four cases: two reputed accounts (*"@who_europe"*, *"@dhsc-govuk"*) and two filters (*"coronavirus treatment"*, *"alternative medicine coronavirus"*). One can expect that the two accounts publish more reliable contents, while the tweets associated to the filters include more dubious information. The sample used to run this demonstration was extracted during the first lockdown period (May 2020) over the full Twitter stream.

The dashboard consists of an upper part with configurable elements, general statistics, keywords extracted from the tweets (computed using *TF/IDF*) and an initial graph that plots the number of tweets (see Figure 5.2). The user can choose to analyse tweets submitted by an account ("issued" in the interface) or "mentions" to an account or to a given keyword query. For this demo, we pre-configured some example queries and the user can click on them and obtain the corresponding results. The granularity of the analysis is also configurable (days, weeks, months). In this upper part, general count statistics and keywords provide the user with a first glimpse of the account/topic in social media.

The bottom part consists of bar graphs that represent the evolution of the sentiment and the credibility of the posted contents (see Figure 5.3). Using this tool, one can observe, for example, that *@who_europe* tends to publish more credible contents (as estimated by the classifier) that the contents associated to tweets that mention words like *"coronavirus treatment"*.

It might be surprising that some contents from a reputed organisation such as the WHO are classified as "highly uncredible". This may be due to false negatives in our predictive technology, which has still room for improvement. However, as mentioned above, the tool identifies general trends and, in general, is able to distinguish the relative quality of authoritative accounts versus more dubious contents (e.g.,"alternate medicine coronavirus").

5.1.4 Final remarks

In this section, we presented an end-user oriented tool called Social Minder. It allows monitoring Twitter but it could be expandable to other web sources, and it provides different estimations (like sentiment or credibility) that can be useful for commercial or research purposes, like monitoring a company's account or analysing misinformation trends.

The demo focuses on one possible use case, but this technology could be adapted to monitor new dynamic text streams, new queries, and/or add new modules, just to name a few.

[1] `http://tec.citius.usc.es/social-minder/`

5.2 AN UNSUPERVISED PERPLEXITY-BASED APPROACH FOR BOILERPLATE REMOVAL

5.2.1 Background

Seminal approaches to boilerplate removal were rule-based. For example, [59] utilised the position of HTML tags in web pages to determine a series of heuristics for extracting the main content. However, this method is rather rigid and cannot handle web documents that follow new structural patterns.

Later, some research teams considered DOM trees and HTML pages divided into blocks and employed supervised learning methods to discern between useful and non-useful blocks. Bauer and colleagues [13] used Support Vector Machines (SVMs) to classify blocks based on linguistic, visual and structural features. Spousta et al [156] proposed a cleaning tool that follows a sequence-labelling approach based on Conditional Random Fields. Kohlschütter and colleagues [93] performed a comparison between shallow textual features and more sophisticated approaches. This comparative study was run with SVMs and under a block-oriented strategy. Pomikálek [132] also presented an unsupervised approach (named *jusText*) to deal with boilerplate based on hyperparameter exploration. This was the first algorithm that took context into account when identifying blocks of text.

A recent proposal is Web2Text [164], which introduces a set of features from adjacent neighbours in the DOM tree and utilises deep learning techniques to predict each block's category. Leonhardt and colleagues [97] noticed that these models require a large number of hand-crafted features and annotated training data. For this reason, they offered an alternative model that does not require pre-processing and directly classifies sequences of raw HTML from few training examples.

Most of these studies have shown the usefulness of their methods in tasks such as search, where pre-processing had been classically confined to simple strategies such as stemming or stopword removal [91]. In this study, we try to go one step further by evaluating the effectiveness and efficiency of our boilerplate removal method not only for a search task but also for document classification and cleaning tasks.

The main novelty of this chapter lies in the utilisation of perplexity as an indicator of well-formed text. Thus, our method can not only remove unnecessary scrapped HTML blocks, but also malformed content. Instead of block segmentation, we propose a more general sentence segmentation technique. We build a Language Model and employ perplexity to estimate sentence likelihood (see Section 5.2.2 for more details). Related to our work, Wenzek et al. [171] employed perplexity as a proxy of quality of documents and scored documents to create curated monolingual corpora. We are not interested in removing documents from the collections but, rather, we define a document pre-processing technique able to remove noisy parts of the original texts.

Other uses of perplexity-based metrics have been suggested for some language-related tasks, such as tweet classification [73], language distance [68, 28] or misinformation identification [96]. Wu et al [173] also employed perplexity scoring as an evaluation measure to estimate the quality of their entity extraction model. Solorio and colleagues [154] used NLP techniques to determine linguistic profiles in children, also using an LM background and the role of perplexity as an assessment metric.

As a final note, observe that we adopted in our study perplexity as our main indicator device, but other language metrics could have been considered. For example, multiple language divergence measures, such as Kullback-Leibler divergence (KLD) [48], could be applied to this task. The exploration of other divergence-based metrics and the study of their connection with the perplexity-based approach reported here is left for future work.

5.2.2 Methodology

Perplexity is a measure that has been mainly employed to evaluate LMs without targeting a specific downstream task [151] (i.e., as an intrinsic evaluation of models of language). A perplexity model indicates how well the data fits into the model distribution. If we assume that the model distribution is correct and unbiased, perplexity allows us to identify noisy data and outliers. In our case, we adopt perplexity as an indicator of potential boilerplate within a webpage.

The perplexity score estimates how well a language model fits a text sample, for instance a word or a sentence. Low perplexity suggests that the language model is good at predicting a given word, sentence or textual extract, while high perplexity indicates that the language model is not good for that prediction. More formally, the perplexity (called *Perpl* for short) of a language model on a text sample (e.g., a sentence) is the exponential of the cross entropy of the given text. Given a sentence $S = w_1, w_2, ..., w_n$ and a language model LM with n-gram probabilities, $P(\cdot)$, estimated on a large corpus, the *Perpl* of S given the n-gram model LM is computed as follows:

$$Perpl(S, LM) = 2^{-\frac{1}{n} \cdot \sum_i^n \log_2 P(w_i | w_1^{i-1})} \tag{5.1}$$

where n-gram probabilities $P(\cdot)$ of a word in position i given the immediate sequence to the left are defined as:

$$P(w_i | w_1^{i-1}) = \frac{C(w_1^{i-1} w_i)}{C(w_1^{i-1})} \tag{5.2}$$

Equation 5.2 estimates the n-gram probability by dividing the observed frequency (C) of a particular sequence of words by the observed frequency of the same sequence without the last word.

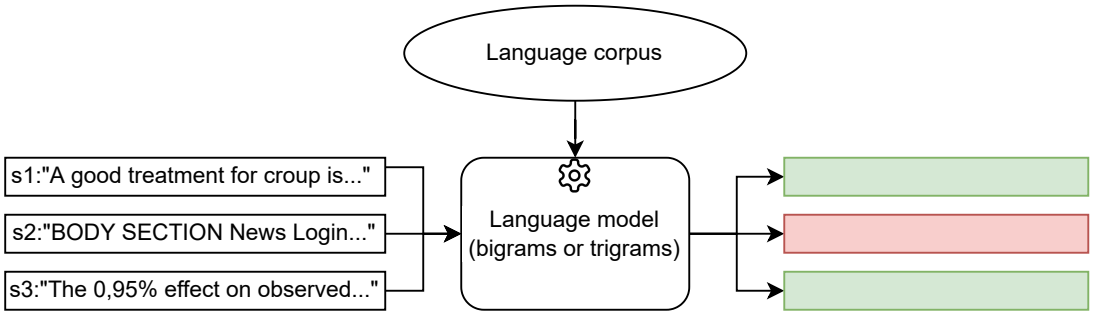


Figure 5.4: Perplexity model for boilerplate removal.

The use of perplexity for cleaning and pruning a corpus is illustrated in Figure 5.4. In our method, perplexity is computed on all input sentences and those with *Perpl* scores above a certain threshold (e.g., > 1000) are removed. This represents a simple but potentially effective way to remove noisy parts of webpages. Traditional pruning mechanisms (e.g., word-level techniques such as stopword removal, frequency-based word pruning, or stemming and lemmatization) are also rather straightforward but, over the years, have become standard elements in multiple text mining and retrieval tools. Note also that the removal of noisy and non-relevant contents can be beneficial not only in terms of effectiveness but also in terms of efficiency (e.g., lighter indexes or data structures, faster access times).

5.2.3 Perplexity models

Two different datasets were selected to build the Language Models: the CORD-19 dataset [167] and the British National Corpus (BNC) [46]. The first is a language corpus associated to a specific topic, while the second represents a more general use of language. CORD-19 contains 50K chapters with over 41K full texts about COVID-19 and related historical coronaviruses such as MERS or SARS. We employed the abstracts of the chapters to build the LM since abstracts are well-formed and contain succinct sentences. These abstracts add up 510 MB of text. BNC is a 100 million word collection of written and spoken British English collected from several sources and created by Oxford University. Work on the corpus began in 1991 and the written part, which is the one that interests us the most here, represents 90% of the total, including extracts from newschapters, specialist journals, academic books, popular fiction, etc. In total, we worked with 620 MB of data to build the LM.

We preferred highly curated texts to create the LMs and we opted for two collections from different domains and genres, one more specific and another one more general. This helps to study the influence of specificity vs generality on perplexity-based pruning of webpages. For this task, there was no need for an intricate model, since the objective is to just discern between well-formed sentences and notorious boilerplate. We chose simple bigrams and trigrams probabilistic

Table 5.1: General statistics of search collections.

task	# queries	query ids	avg # rel docs per query	used for	doc collection	# docs	type of docs
TREC 2019 Decision Track	51	1–51	82	Hyperparameter selection	ClueWeb12-B13	50M	webpages
TREC 2013 Web Track	50	201–250	88	Test	ClueWeb12-B13	50M	webpages

models, which are both fast and adequate for perplexity computation.

5.2.4 Experimental settings

5.2.4.1 Datasets

The perplexity models built from CORD-19 and BNC were evaluated with two test collections designed for specific tasks (see Table 5.1 for details). The ClueWeb12-B13[2] collection was employed to perform experiments of indexing and search. This collection contains approximately 50 million pages. We ran tests with two different sets of queries and relevance assessments (one set of search topics –from the medical domain– obtained from the TREC 2019 Decision Track [1] and another set of general topics from the TREC 2013 Web Track [44]). The first set was used to determine the only parameter that our method requires: the maximum admissible value of perplexity per sentence. This parameter selection was then validated with the second set of test queries. The results are reported in Section 5.2.5.

A second class of experiments was performed to assess the effectiveness of the perplexity-based approach within a text classification task (webpage classification). To that end, the WebKB collection[3] was chosen. It contains 8,282 web pages from four different universities classified into seven different categories (courses, departments, faculty, projects, staff, students and other).

Finally, a third set of experiments was conducted to test the ability of our solution to clean up web content in the context of a well-known competition for cleaning webpages. More specifically, the shared-task CleanEval [10] provided the ideal framework for these last experiments. The CleanEval dataset contains a random sample of web corpora which was collected by making queries to Google (only html pages were collected). We employed 625 CleanEval webpages written in English. The collection contains the original webpages and webpages cleaned by recruiting human annotators who read the webpages and removed noisy content.

[2] http://lemurproject.org/clueweb12/
[3] http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

5.2.4.2 Metrics

For testing retrieval performance, the following metrics were used:

– Precision at n ($P@n$): represents the fraction of the top n documents of the ranking that are relevant to the user's information need [47]:

$$P@n = \frac{\sum_{p=1}^{n} rel(p)}{n} \tag{5.3}$$

where $rel(p)$ equals 1 if the item at position p is a relevant document, and 0 otherwise.

– Average Precision (AP): it summarises the ranking by averaging the precision scores at the rank positions where a relevant document is found[4] [47]. In Equation 5.4, $rels$ is the total number of relevant documents and $P@n$ is the precision at a cutoff n. The mean AP (MAP) for a set of queries is the mean of average precision scores for each query. It is a popular method to assess the effectiveness of a ranked set of results.

$$AP = \frac{1}{rels} \cdot \sum_{n=1}^{rels} P@n \cdot rel(n) \tag{5.4}$$

– Normalised Discounted Cumulative Gain at n ($NDCG@n$): $NDCG$ measures the quality of search results taking into account different grades or levels of relevance [47]:

$$NDCG = \frac{DCG}{IDCG} \tag{5.5}$$

$$DCG = \sum_{p=1}^{n} \frac{2^{\mathrm{rel}_p} - 1}{log_2(p+1)} \tag{5.6}$$

DCG, Discounted Cumulative Gain (Equation 5.6), defines the user's gain as a measure that grows as he/she goes from the top of the ranking to lower positions. Under $NDCG$, the gain produced by each ranked document depends on its position. Gains obtained by relevant documents at higher positions are greater than those from relevant documents at lower ones. For example, a highly relevant document at a low-rank position is substantially penalised. To that end, each gain is penalised by a discounting factor ($log_2(p+1)$). The DCG values are normalised by dividing the DCG scores by the ideal DCG ($IDCG$, which represents the gains obtained by an oracle system that ranks documents by decreasing order of their actual relevance). The $NDCG@n$ score represents the accumulated gain that the user obtained from examining the top n results.

[4] and relevant documents that were not retrieved contribute with a 0 to the AP score (i.e. their $P@n$ is set to 0).

All these metrics are computed for each available query and the reported figures represent the mean value across all queries. To analyse the statistical significance of the performance difference between two results, we employ the Wilcoxon test on the paired values (one from each query). Parapar et al. [126, 127] showed that the Wilcoxon test is a highly reliable test to compare retrieval systems (yields more statistical power and fewer type I errors). The significance tests help to determine whether or not each observed difference is anecdotal.

For evaluating classification performance, the F1 score (harmonic mean between precision and recall) for each class and macro average F1 (unweighted mean of F1 per class) are reported.

In this chapter, we also aim at improving computational efficiency and, thus, we report here some time measures, storage improvements, and an estimate of carbon savings [95]. The elimination of noisy document's parts has a potential to reduce time or space complexity, and it is important to quantify these improvements.

For the CleanEval shared-task, the effectiveness metric measures the similarity between a cleaned version of the file (produced by a given cleaning algorithm, e.g. our perplexity-based method) and the gold standard (produced from human annotators). To that end, the task considers Levenshtein edit distance as the main scoring method. It computes the distance between two strings given the fewest operations required to transform one into the other. The final measure is the percentage of misalignment between files (in our case, considering only text; HTML labels are skipped).

5.2.4.3 Search and Classification Models

For the retrieval experiments, the collection was indexed using the Anserini search engine [176], and the retrieval model utilised was the query likelihood model (QL) [133]. This is a well-known probabilistic retrieval approach that assumes that the query is generated by sampling words from the document and that each query term is independent, meaning that the probability of the query is a product of the probability of each term. Equation 5.7 presents the QL document relevance score, where $p(q|d)$ is the conditional probability of the query q given the document d. To account for query terms that do not appear in the document, probability values need to be smoothed using a reference language model or corpus. In our experiments, we utilised Dirichlet Smoothing (Equation 5.8). A full description of probabilistic language models for IR, which shows the advantages of QL with Dirichlet smoothing and its solid length normalisation abilities, is available at [108].

$$\text{QL\_score}(q,d) = \log p(q|d) \tag{5.7}$$

$$\text{QL\_score}_{\text{DIR}}(q,d) = \sum_{w \in q,d} c(w,q) \cdot \log\left(1 + \frac{c(w,d)}{\mu \cdot p(w|C)}\right) + |q| \cdot \log\frac{\mu}{\mu + |d|} \tag{5.8}$$

$c(w,q)$ ($c(w,d)$) is the number of occurrences of word w in the query (document). $p(w|C)$ is the probability of the word in the background corpus and $|q|$ and $|d|$ are the lengths of the query and document, respectively. μ is the smoothing parameter, which was set to 2000 in our experiments.

For document classification, state-of-the-art Transformers models were used [161]. More specifically, we evaluated and compared BERT [50] and RoBERTa [106] models. These models consist only of an encoder architecture and have proven ability in outperforming traditional classifiers. We experimented with the base and large configurations of both models. To that end, Ernie[5], the open-source library, was utilised.

Following [157], we did cross-validation by always leaving the webpages from one of the universities out. The classification collection is highly imbalanced and, thus, we followed standard practise to set the error weights from the proportion of cases in each class.

5.2.4.4 Hardware platform

Different hardware configurations were used, depending on the task. The tag removal and perplexity computation for the ClueWeb dataset (which corresponds to *TREC 2019 Decision Track* and *TREC 2013 Web Track* tasks) was carried out in a Big Data computing cluster. This cluster is formed by 15 nodes, each one with two Intel Xeon E5-2630 v4 and 384GB of RAM, for a total of 20 cores per node. The computing load for these tasks was distributed across the 300 available cores using data parallelism. Once the tag removal and perplexity computation for the dataset was completed, indexing and search (supported by Anserini) were run in one node of the cluster.

The classification task required a GPU-powered server. For this reason, experiments were conducted in a single server with two Intel Xeon Gold 5220, 192GB of RAM and two Nvidia Tesla V100S. These GPUs efficiently support the training processes required by BERT and RoBERTa classification models.

Finally, the cleaning task was computed on a single node with an Intel i7-9700K CPU @ 3.60GHz, 32GB of RAM and no GPU capabilities.

5.2.5 Results

5.2.5.1 Search experiments

TREC 2019 Decision Track

The first part of the experiments was conducted with queries and relevance assessments from the TREC 2019 Decision Track (DT19). This test collection consists of search topics related to

[5] https://github.com/labteral/ernie

	P@5	P@10	AP	NDCG @10	Size (GB)
BASELINE	0.564	0.546	0.328	0.458	382
Bigrams model					
CORD-19 (1K)	0.592	0.540	0.321	0.484	51
CORD-19 (2K)	0.592	0.550	0.332	0.479	61
CORD-19 (4K)	0.608	0.564	0.335	0.489	72
CORD-19 (6K)	0.596	0.566	0.335	0.489	78
CORD-19 (8K)	0.584	0.556	0.326	0.477	82
CORD-19 (10K)	0.584	0.552	0.325	0.468	85
CORD-19 (12K)	0.580	0.562	0.323	0.473	88
CORD-19 (14K)	0.588	0.568	0.326	0.477	90
BNC (1K)	0.620	0.554	0.314	0.484	60
BNC (2K)	0.608	0.580	0.333	0.492	69
BNC (4K)	0.604	0.578	0.338	0.487	78
BNC (6K)	0.620	0.578	0.337	0.487↑	89
BNC (8K)	0.600	0.568	0.334	0.481	93
BNC (10K)	0.600	0.576	0.333	0.486↑	96
BNC (12K)	0.604	0.576	0.333	0.486	99
BNC (14K)	0.600	0.566	0.335	0.481	101
Trigrams model					
CORD-19 (1K)	0.564	0.526	0.293↓	0.455	43
CORD-19 (2K)	0.612	0.552	0.328	0.485	55
CORD-19 (4K)	0.604	0.576	0.336	0.493	66
CORD-19 (6K)	0.604↑	0.564	0.334	0.488	72
CORD-19 (8K)	0.608↑	0.564	0.338	0.487	77
CORD-19 (10K)	0.608	0.572	0.336	0.490↑	80
CORD-19 (12K)	0.592	0.572	0.332	0.485	83
CORD-19 (14K)	0.596	0.566	0.330	0.480	86
BNC (1K)	0.584	0.546	0.293↓	0.455	52
BNC (2K)	0.628 ↑	0.560	0.324	0.490	63
BNC (4K)	0.620↑	0.582	0.335	0.494	72
BNC (6K)	0.620↑	0.580	0.338	0.491	77
BNC (8K)	0.624↑	0.576	0.341	0.486	81
BNC (10K)	0.620↑	0.578	0.340	0.490↑	84
BNC (12K)	0.604	0.580	0.339	0.491↑	87
BNC (14K)	0.596	0.578	0.335	0.486↑	89

Table 5.2: TREC 2019 Decision Track. Effect of different perplexity cutoff values (reported in brackets) on retrieval performance. The ↑/↓ symbols indicate whether the method significantly improves or not (Wilcoxon test, $\alpha = 0.05$) over the baseline (no perplexity-based removal).

the medical domain and documents crawled from the web (ClueWeb dataset).

The DT19 collection was used for optimising the only hyperparameter of the boilerplate removal method: the perplexity cutoff value. The higher the perplexity cutoff the fewer sentences are removed. We experimented with the following perplexity cutoff values: 1K, 2K, 4K, 6K, 8K, 10K, 12K and 14K.

Table 5.2 shows the results by comparing the effectiveness of different levels of removal against the retrieval baseline, which does not remove any sentence. Most of the tested models and thresholds outperform the baseline and in some cases the improvements are statistically significant. The improvements are solid for high-precision metrics (P@5 and P@10) but also for AP and NDCG, which take into account the entire ranking of documents. Only a couple of instances lead to performance decreases that are statistically significant (and this only happens for AP). This is reasonable since AP is a recall-oriented measure. Our cleaning technique eliminates malformed text but the method is not perfect and, thus, some pruned sentences might be actually

relevant. This affects AP in some of the tested configurations. In contrast, the removal method proves to be powerful for the three precision-oriented metrics (P@5, P@10 and NDCG@10).

This exploration of the perplexity cutoff suggests that we can safely find a cutoff configuration that is robust and works well for precision-oriented and recall-oriented metrics. The reader should also bear in mind that it is important to keep (or even improve) retrieval effectiveness, but space savings and other efficiency-oriented factors are other important aspects associated to perplexity-based removal (this will be further discussed in Section 5.2.5.1).

Not surprisingly, BNC-based models outperform CORD-19 ones. This confirms that the BNC-based approach, which removes noisy sentences based on a general model of language, is more apt to determine which sentences should remain. The DT19 search topics are medical queries and, thus, somehow close to CORD-19 but, still, this more focused language data does not give any added value. Observe that it is convenient that the approach does not require a topic-specific corpus to build the perplexity models. A general language corpus suffices to filter out noisy and off-topic contents and the BNC-based method would be applicable across general-purpose search tasks and collections.

Regarding bigrams vs trigrams, the latter models yielded slightly better effectiveness. Taking these results into account, we adopted the following models, which constitute a good balance among all the performance metrics: BNC trigram models (8K and 10K), and BNC bigram models (8K and 10K). These variants, which are marked in light grey color in the table, are not always the best performers but they represent a solid configuration that often leads to statistical significant improvements.

TREC 2013 Web Track

Next, the selected models were evaluated with another test collection, the TREC 2013 Web Track (WT2013). This dataset has also a large corpus of webpages crawled from the web (ClueWeb12 dataset) and includes search topics that represent general information needs (informational or navigational queries).

The results of this experimentation are shown in Table 5.3. In most of the cases, the perplexity-based variants lead to higher effectiveness and the method only yields minor decreases in AP. Note also that AP is not a crucial metric for most web retrieval tasks, where high recall is rarely pursued. This outcome further reinforces the potential of the approach not only to produce lighter and less noisy indexed data, but also to maintain and even improve retrieval performance.

Additionally, there is not a noticeable difference between bigram and trigram models. Both alternatives perform roughly the same. This is another interesting outcome since bigram models are simpler and thus less costly.

	P@5	P@10	AP	NDCG @10	Size (GB)
BASELINE	0.236	0.230	0.039	0.154	382
Bigrams model					
BNC (8K)	0.268	0.230	0.037	0.163	93
BNC (10K)	0.264	0.248	0.037	0.165	96
Trigrams model					
BNC (8K)	0.264	0.238	0.037	0.171	81
BNC (10K)	0.268	0.230	0.037	0.165	84

Table 5.3: TREC 2013 Web Track. The ↑/↓ symbols indicate whether the method significantly improves or not (Wilcoxon test, $\alpha = 0.05$) over the baseline (no perplexity-based removal).

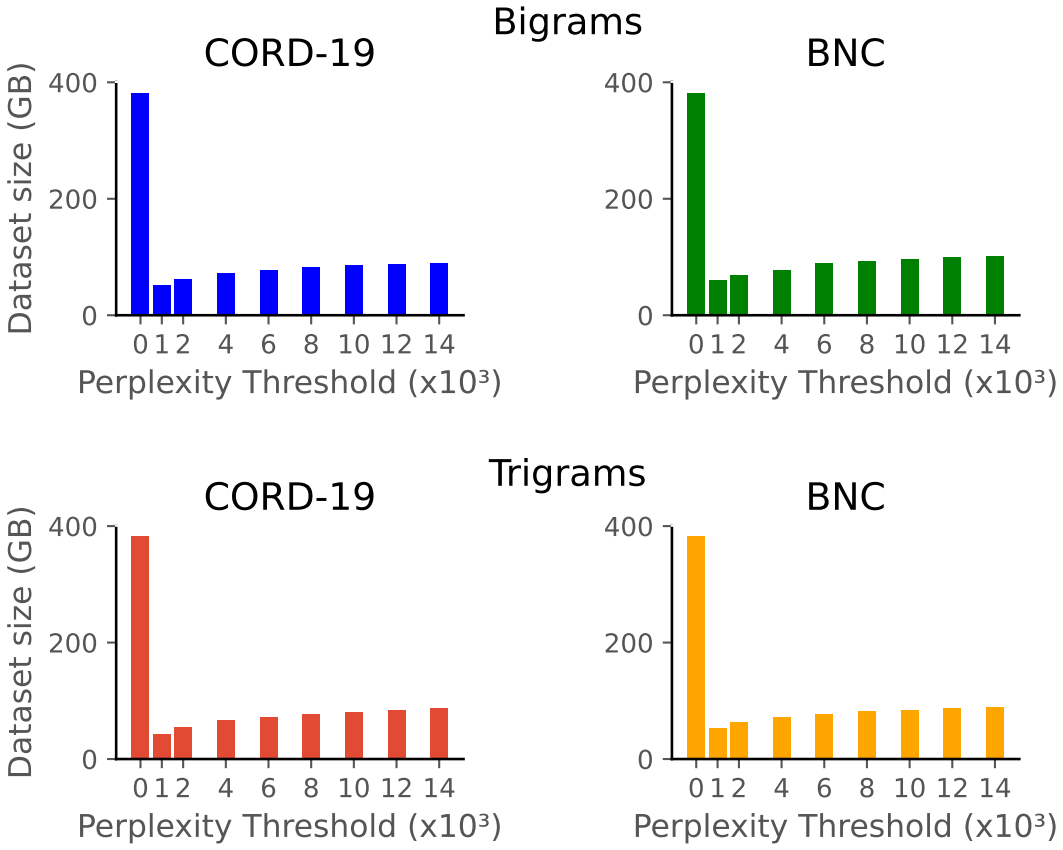


Figure 5.5: Space savings derived from using our model in the IR search task.

Efficiency

Improving effectiveness is an attractive feature of the perplexity-based removal but efficiency is also a major dimension that needs to be taken into account. The first notable advantage of our method is space saving. The sizes of the corresponding indexes are reported in Tables 5.2 and 5.3 (last column), while Figure 5.5 plots the sizes in a graphical way (the leftmost bar represents the baseline situation with no perplexity-based removal). Compared with the original collection, we can save up more than 75% of the storage cost of the indexing structures (e.g., 96 GB vs 382 GB for the BNC-10k model).

	Tag cleaning	Perplexity computing	Index building	Search (50 queries)	Search (10K queries)
BASELINE	-	-	03:24:23	00:00:45	125:25:00
Bigrams model					
BNC (8K)	00:45:00	00:45:00	01:01:42	00:00:23	64:11:40
BNC (10K)	00:45:00	00:45:00	01:07:48	00:00:25	70:03:20
Trigrams model					
BNC (8K)	00:45:00	00:45:00	01:00:08	00:00:24	65:31:40
BNC (10K)	00:45:00	00:45:00	01:04:04	00:00:25	68:56:40

Table 5.4: Time measurements (HH:MM:SS) for the baseline and some perplexity-based variants.

This space saving also translates into several time improvements. The reduced size of the collection results in a substantial decrease of the computing time required to index and search (see Table 5.4). Our method requires two cleaning phases before indexing: the removal of HTML tags from the entire collection and the actual computation and removal of perplexity at sentence level. As stated in Section 5.2.4.4, both operations were carried out in parallel in a Big Data cluster. However, with multiple users doing search online, a crucial measure for an IR engine is search time and that is where our method stands out. For example, for the BNC bigrams model with a threshold set at 8K, the search of 10,000 queries is approximately 1.95 times faster than the same search against the original index. Indexing and searching were executed on a single node because of the restrictions of the retrieval library utilised. Nonetheless, these results can generalise to an entire cluster.

Consequently, this translates into important carbon emission reductions. This reduction grows with the number of searches done against the collection. In Table 5.5, we show the estimated carbon emissions for our system. Carbon emission were estimated following the methodology presented in [95], while carbon intensity data for our region and energy provider was taken from [119]. We first calculated the equivalent CPU-hours required for the computation in each step and then derived the estimated carbon emissions with the following relation:

$$\text{eq. kg of } CO_2 = \frac{t \cdot C_e \cdot W_{\text{cpu}}}{1000} \tag{5.9}$$

where t is the equivalent CPU-hours of computation (taking parallelization into account), $C_e = 0,341$ is the carbon efficiency coefficient of the grid (measured in kg CO_2eq/kWh) and W_{cpu} is the Thermal Design Power of the CPU in watts.

With the 8K BNC-bigrams model, the answer to 10,000 searches requires approximately half carbon emissions of the baseline. Consider, for example, the case of a real-world IR system. Google reported more than 2 million searches per minute in 2012, its latest statistics available[6]. Taking all the perplexity computing, indexing and inference emissions into account, our method reduces on average 5.58kg of CO_2 per million queries, which translates into more than 22 tons of CO_2 saved each day.

[6] https://archive.google.com/zeitgeist/2012/#the-world

	Tag cleaning	Perplexity computing	Index building	Search (10K queries)	Search (1M queries)
BASELINE	–	–	0.20	0.15	14.54
Bigrams model					
BNC (8K)	0.65	0.65	0.06	0.07	7.44
BNC (10K)	0.65	0.65	0.07	0.08	8.12
Trigrams model					
BNC (8K)	0.65	0.65	0.06	0.08	7.60
BNC (10K)	0.65	0.65	0.06	0.08	7.99

Table 5.5: Carbon emissions in kg of CO_2 for the baseline and some perplexity-based variants.

	F1 Macro	F1 Course	F1 Department	F1 Faculty	F1 Project	F1 Staff	F1 Student	F1 Other
BERT base								
BASE. CLEAN	0.5818	0.6765	0.0942	0.7676	0.3914	0.4594	0.8160	0.8678
Bigr. BNC (8K)	**0.5832**	0.6407	0.1083	0.7555	0.4327	0.5425	0.7492	0.8534
Bigr. BNC (10K)	0.5667	0.5592	0.1278	0.7538	0.4123	0.4806	0.7943	0.8391
Trigr. BNC (8K)	0.5128	0.5217	0.0294	0.7064	0.4201	0.3564	0.7221	0.8332
Trigr. BNC (10K)	0.5280	0.5358	0.1377	0.7001	0.3881	0.4058	0.7098	0.8188
RoBERTa base								
BASE. CLEAN	0.6109	0.6948	0.2190	0.7589	0.4183	0.4753	0.8230	0.8875
Bigr. BNC (8K)	**0.6350**	0.6980	0.2917	0.7553	0.5444	0.5425	0.7957	0.8911
Bigr. BNC (10K)	0.6137	0.6781	0.1881	0.7681	0.4422	0.5497	0.7871	0.8828
Trigr. BNC (8K)	0.5928	0.6293	0.2399	0.7523	0.4301	0.4739	0.7493	0.8805
Trigr. BNC (10K)	0.5544	0.5928	0.1210	0.7571	0.4182	0.4555	0.7203	0.8232
BERT large								
BASE. CLEAN	0.4309	0.4749	0.2131	0.5979	0.2332	0.2623	0.6194	0.6155
Bigr. BNC (8K)	**0.5815**	0.6454	0.1372	0.7806	0.4568	0.4111	0.7564	0.8831
Bigr. BNC (10K)	0.5211	0.5313	0.1165	0.7368	0.3600	0.3978	0.7555	0.7502
Trigr. BNC (8K)	0.5419	0.5387	0.1061	0.7444	0.4837	0.3339	0.7445	0.8418
Trigr. BNC (10K)	0.5320	0.5576	0.0583	0.7244	0.3916	0.4340	0.7469	0.8108
RoBERTa large								
BASE. CLEAN	0.6255	0.7406	0.1929	0.7809	0.4148	0.5357	0.8157	0.8980
Bigr. BNC (8K)	0.6169	0.6632	0.1387	0.7490	0.4964	0.5476	0.8227	0.9013
Bigr. BNC (10K)	0.6282	0.7198	0.2137	0.7955	0.4478	0.4701	0.8416	0.9050
Trigr. BNC (8K)	0.5297	0.6355	0.0888	0.7315	0.3422	0.3588	0.7433	0.8082
Trigr. BNC (10K)	**0.6289**	0.6053	0.4125	0.7961	0.4150	0.5468	0.7592	0.8677

Table 5.6: Classification results for WebKB dataset. For each block, the top performer (F1 macro) is bolded.

5.2.5.2 Classification experiments: WebKB

Let us evaluate now the performance of the perplexity-based models for another text-related challenge, a document classification task. To that goal, we adopted the experimental methodology and collection utilized in [157], but we worked with newer classifiers based on transformers (instead of traditional SVMs).

The results are shown in Table 5.6. The first row corresponds with the baseline that, in this case, takes the webpages and only removes the HTML tags. This is a standard pre-processing approach in web classification. Our cleaning methods allowed the classifier to perform better in many cases. The bigrams models (and, particularly, the ones with the threshold set to 8K) are the best performers. Another interesting outcome is the low performance obtained by the baseline with the BERT large model. A plausible explanation for this is that BERT large is a model with a huge number of parameters, and it would need a larger amount of training data

	Effectiveness (%)
jusText	76.20
Bigrams model	
BNC (8K)	79.76
BNC (10K)	79.80

Table 5.7: CleanEval shared-task results for two perplexity-based methods and the jusText algorithm. The percentage scores represent the average similarity between the webpages cleaned automatically and the ground truth webpages.

to generalise better. In any case, the BERT large model also benefits from our pre-processing methods, which only keep useful information and remove boilerplate. This difference is not as noticeable with RoBERTa large, as it is an improved and robust model to avoid such problems.

Overall, these results suggest that a base model with perplexity-based removal of noisy sentences is comparable to (or better than) more sophisticated (and more inefficient) models based on a larger set of parameters. As a matter of fact, the highest F1 macro (0.6350) is obtained by combining the RoBERTa base model with perplexity-based pre-processing (bigrams, 8k threshold).

Efficiency

In terms of efficiency, there is again a saving in space by reducing the size of the dataset. However, for this task the space reduction is less significant as it is a collection of a few megabytes. As stated before, our boilerplate removal method permits the utilisation of less expensive models, such as RoBERTa base which, together with the perplexity-based pre-processing, overcomes more complex models. This results in direct savings in training time. For instance, in our experiments the RoBERTa base model was more than 3 times faster than RoBERTa large.

On the other hand, perplexity-based pruning leads to smaller representations of the webpages and, potentially, to lower prediction times. However, the state-of-the-art classification models described above, based on transformer technologies, do not analyse the entire input documents but are limited to a 512 token limit. In practice, even after removal of noisy sentences, most documents are above the 512 token limit and, thus, the improvement in prediction time was not noticeable. However, the decrease in prediction time would be observable under other classification models that make predictions based on the whole page.

5.2.5.3 Cleaning experiments: CleanEval

As stated before, our method goes one step further than simply removing boilerplate. Its main goal is to identify useful and well-formed text to enhance the performance of downstream tasks. However, we also wanted to demonstrate its validity for cleaning webpages. To this end, we selected the best performers for the previous tasks (both BNC-based bigrams models with thresh-

olds set to 8k and 10k) and compared them with the jusText algorithm. As can be seen in Table 5.7, our models outperform jusText for the CleanEval shared-task.

5.2.6 Discussion

Our unsupervised cleaning method improved performance for most of the evaluation measures under the search task. However, it should be noticed that the perplexity-based removal works as a precision-oriented technique, and it was less effective in terms of recall (particularly in terms of AP). This occurs because our cleaning technique eliminates not only boilerplate, but also malformed or noisy sentences and some of them might be actually relevant.

Another important outcome is that for both tasks BNC-based models outperformed their CORD-19 counterparts. We wanted to test how the topicality of the background corpus influences performance. The results suggest that the most general model, in terms of content, yields the best results. This happened for both tasks (search and classification).

The proposed perplexity models consisted of simple bigrams and trigrams under a probabilistic approach. Our experiments have demonstrated that the bigram variants, which are less complex and thus more efficient, are also the top performers for most of the evaluated metrics. This saving was also one of the main objectives of the study, as stated at the beginning.

In our study, not only effectiveness but also efficiency was taken into account. As detailed in Section 5.2.5.1, our models substantially reduce carbon emissions, when a sufficient number of searches are performed over the collection [95]. We estimate that our method could reduce in average 5.58kg of CO_2 per million queries. To put this result in perspective, this CO_2 weight is equivalent to 24.13 km driven by an average car[7], 3.09 kg of coal burned [8] or 0.1 tree seedlings consuming carbon for 10 years[9].

Finally, we have done a special effort to make the perplexity-based technology available to the community. We anticipate multiple uses of perplexity as a pre-processing mechanism in a wide range of text mining projects. To facilitate new applications of perplexity, we have created a Python package and a web demo that implement the boilerplate removal methods detailed in this chapter. This is described in the next section.

5.2.7 Python package and Web demo

We created and released PyPlexity[10], a Python package available at the PyPi[11] repository. The package includes the source code required to train and utilise the models. This library serves two

[7] Gases equivalence calculator for driven miles
[8] Gases equivalence calculator for burned coal
[9] Gases equivalence calculator for tree seedlings consuming carbon for 10 years
[10] `https://github.com/citiususc/pyplexity`
[11] `https://pypi.org/project/pyplexity/`

purposes: on one hand, allows end-users to utilise our models from any python program, with just a single line of code. On the other hand, it offers a command-line interface that provides straightforward perplexity computation.

There are three main commands that can be run from console. The simplest one is *perplexity*, which calculates the perplexity score for a sentence and a given model. The *tag-remover* command processes a raw HTML file or input directory of files and removes any tags and other non-textual components. Finally, the *bulk-perplexity* processor computes perplexity for a batch of files and removes any sentence above the perplexity limit provided by the user (set by default to 8K). These two latter commands also have distributed computing capabilities, allowing efficient processing of large collections of documents in a computing cluster[12].

The Python interface allows integrating these computations into Python code. To demonstrate some of these capabilities, Figure 5.6 shows how to utilise our tool to compute the perplexity score of a sentence and clean a line of text from Python code. Please refer to the documentation[13] for further details.

```
[1] from pyplexity import PerplexityModel

    model = PerplexityModel.from_str("bigrams-cord19")
    perpl = model.compute_sentence("this is normal text")
    print(perpl)

    downloading: 100%|##########| 233M/233M [00:12<00:00, 19.4MiB/s]Loading model...
    Done.
    5375.94107162351

[4] from pyplexity import PerplexityProcessor

    text_processor = PerplexityProcessor(perpl_model=model, perpl_limit=7000.0)
    clean_text = text_processor.process("This is a normal sentence. Meanwhile,
        hjldfuia HTML BODY this one will be deleted LINK URL COUISUDOANLHJWQKEJK")
    print(clean_text)

     This is a normal sentence.
```

Figure 5.6: Some of the Pyplexity capabilities shown up and running in a Jupyter Notebook.

Additionally, we published a web app[14] that permits users to test the basic functionalities of the library. It allows to directly clean a raw text or html input or to clean a webpage identified by its URL. A screenshot of the demo is shown in Figure 5.7.

5.2.8 Final remarks

In this section, we have proposed an unsupervised method for extracting useful content from scrapped webpages. To the best of our knowledge, this is the first perplexity-based approach adopted for this kind of text pre-preprocessing.

[12] https://github.com/citiususc/pyplexity#parallel-mode-cluster
[13] https://github.com/citiususc/pyplexity#interfacing-from-python
[14] https://tec.citius.usc.es/pyplexity/

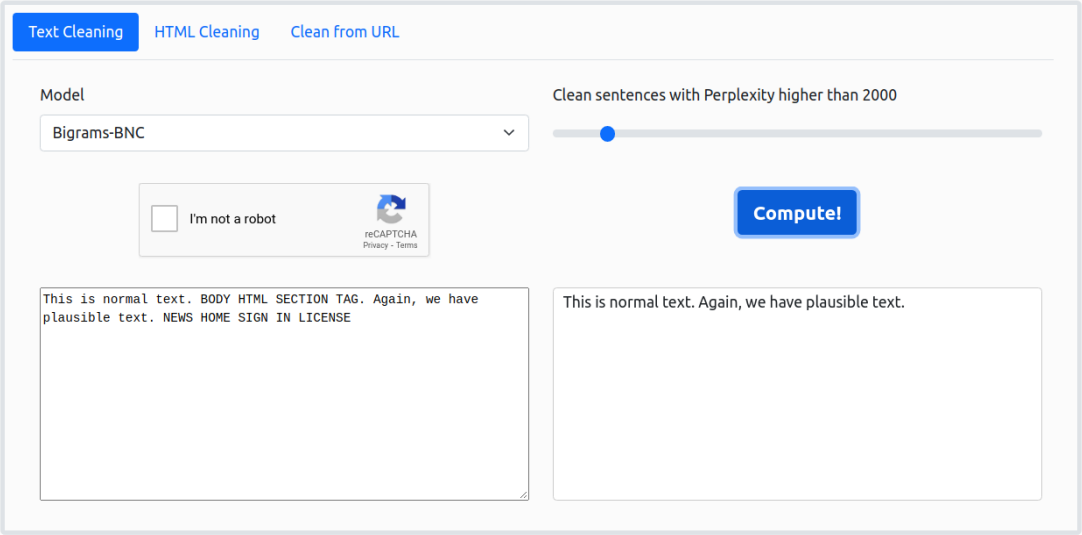


Figure 5.7: Demo webpage that showcases the capabilities of the cleaning tool presented in this section.

A thorough experimentation process has been performed. This included evaluation with different test collections and tasks. With perplexity-based removal of sentences, we have managed to improve performance under several web-related tasks, in some cases even significantly.

In addition, our models also follow Green AI principles and seek to maximise energy efficiency. The pre-processing requirements of our method are compensated for the advantages associated with the resulting noise-free web contents.

Furthermore, we provide an easy-to-use library and web tool, which facilitate the adoption of this technology to support the removal of noisy textual extracts. We are convinced that perplexity-based removal might play a role in multiple application domains and text mining tasks.

As future work, we intend to explore other tokenisation techniques and work with different textual granularities (besides sentence level) to see the impact on recall-oriented measures. Moreover, we also plan to test the helpfulness of these cleaning techniques in other areas, and continue expanding the toolkit of open-source tools.

6 Conclusions

The advent of digital media has facilitated access to information [145], yet the results provided are not always reliable [2], accurate [52], or high-quality [146]. Misinformation detection is a research challenge that has gained traction in the last decade.

In particular, health misinformation detection has been a focal point for the scientific community, attracting numerous research efforts [163, 155, 2]. The urgent need for effective misinformation detection tools was especially evident during the 2020 pandemic when a substantial amount of COVID-19-related information was questionable or of poor quality [84, 129]. In this context, the early identification of health-related unreliable information is crucial to preventing potential harm [162].

This necessity has driven the definition of shared-tasks and the creation of new test collections to foster research on health misinformation and encourage the dissemination of correct and credible information. For example, the CLEF eHealth initiative [86] and the TREC HM Track [42, 41] had an intense activity within the last decade. These campaigns have functioned as a testing ground for assessing the reliability of our proposed solutions.

In this thesis we addressed the creation and evaluation of automatic classification and search methods to support end-users in their access to reliable health-related contents. In chapter 2, we presented our first attempts to approach the problem of health misinformation detection. We replicated a seminal study oriented to reliability classification of medical websites, and we compared traditional machine learning solutions against new deep linguistic models for this task. The main conclusions derived from this initial work were the following:

- We confirmed the main trends derived from the original study. In our experiments, models built from word-based features or models that combine multiple types of features stand out from competing alternatives.

- We also tested our predictive technology against two new test sets, and the conclusions remained the same. The technology developed to support these reproducibility experiments was the basis for our participation in TREC 2020 HM Track. However, the classic document-level approach with sparse word features did not generalise well to the TREC

data and we concluded that the learned classifiers might be excessively dependent on the specific topics and expressions employed in the training data.

– By comparing traditional and neural approaches, we also found that traditional models, like Naive Bayes, still represent a consistent approach for some classification tasks. For example, in readability classification, NBs should not be discarded as an alternative computationally lightweight method.

In Chapter 3, we presented our own multistage retrieval system for health-related misinformation detection. We can summarise its main contributions as follows:

– We conducted a thorough comparative study to validate the potential of our platform for a socially worrying matter: detecting COVID-19 misinformation. Our analysis assessed the effectiveness of multiple stages, including document retrieval, passage retrieval, and reliability estimation. We demonstrated that every stage was valuable to improve performance.

– The fusion of multiple forms of evidence led to the most efficient misinformation estimation methods. However, not all solutions were equally good, since some unsupervised fusion solutions were far better than learning-to-rank approaches.

– We also demonstrated that our top performing variants are competitive with state-of-the-art methods.

– The results of our analysis also showed that certain signals or pieces of evidence help more in finding helpful documents, while others focus on limiting the retrieval of harmful contents. However, it is still difficult to find a one-fits-all solution that represents a good trade between both sides of the problem.

– We also participated with this system in the TREC 2021 HM Track, obtaining a meritorious third position.

In Chapter 4, we put under scrutiny the capacity of the new LLMs to provide correct medical advice. To that end, we conducted a thorough evaluation, testing different prompts, and experimenting with zero- and few-shot approaches. This evaluation yielded the following conclusions:

– We demonstrated that the selected prompt strongly influences performance. More specifically, the prompts that bias LLMs towards more reputable sources tend to perform better.

– For some specific models and prompts (the simplest ones), including in-context examples can be very helpful.

– Error analysis on the model's responses showed that, although their effectiveness is remarkable, there are still some concerning mistakes. In some cases, the LLM provides advice against the medical consensus or demonstrates a worrying lack of common sense knowledge.

Finally, in Chapter 5 we presented two complementary tools that aid in online processing of massive textual contents. First, we presented Social Minder, a modular and scalable Big Data platform that can monitor social media publications in real-time. It also provides different estimates, such as those related to sentiment or credibility. The tool was exploited to build a use case oriented towards COVID-19 misinformation monitoring. However, Social Minder can be adapted to monitor new text streams, queries, and to support new functionalities. Second, we have developed a text preprocessing tool named Pyplexity. It consists of a text quality estimator that flags malformed and other low quality sentences within scraped text. This is critical for several NLP tasks. Through our experimentation, we have demonstrated the usefulness of the Pyplexity library, both in terms of efficiency and effectiveness, for several downstream tasks, such as text classification.

Overall, we can conclude that the objectives of the thesis detailed in Section 1.1 have been successfully fulfilled.

6.1 Future work

We can identify several research directions to continue advancing in health misinformation detection:

– One interesting line of research consists of moving towards a user-oriented perspective. We have taken some initial steps in this direction with an study that attempted to define a set of guidelines for a more accurate labelling of highly subjective concepts such as credibility [56]. A strong user-oriented focus is critical for the development of reliable test collections. However, there are still some open research questions. For instance, it has not yet been demonstrated whether the judgements produced with these guidelines correctly represent end-users' perception of credibility. This could be done by conducting a user study in which participants evaluate the credibility of a number of medical websites. In another line, we could also study how users interact with pages of different quality, not only studying the credibility dimension, but also the correctness of the information.

– As we previously stated, information access is evolving from traditional web search towards a more conversational approach, exploiting the abilities of the new LLMs. Following this line of thought, we would like to compare the ability of traditional search engines (SEs) to provide correct medical advice against the responses provided by LLMs. At the

moment, we have some on-going research that involves monitoring the top search results provided by different SEs, such as Yahoo, DuckDuckgo, Bing and Google. The main goal is to determine the effort end users have to make on a SERP to get a correct medical answer and compare that with the effectiveness in information access with generative models.

– We would also like to explore a formal path to use score distribution (SD) models to further understand misinformation detection systems [8, 6, 7]. Given the distinction between harmfulness and helpfulness established in the TREC HM Track, we could, for instance, exploit SDs to determine the optimal threshold at which to stop displaying search engine results with the least possible damage. Another path that we would like to explore is to study the viability of modelling the score distribution of black-box re-ranker models. This approach would shed light towards the complex decision process these models follow.

Bibliography

[1] Mustafa Abualsaud, Christina Lioma, Maria Maistro, Mark D. Smucker, and Guido Zuccon. Overview of the TREC 2019 decision track. In *Proceedings of the 28th Text REtrieval Conference, (TREC '19)*, Gaithesburg, Maryland, USA, 2019. National Institute of Standards and Technology (NIST).

[2] Mustafa Abualsaud and Mark D Smucker. Exposure and order effects of misinformation on health search decisions. In *Proceedings of the 42nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019.

[3] Eugene Agichtein, Eric Brill, Susan Dumais, and Robert Ragno. Learning user interaction models for predicting web search result preferences. In *Proceedings of the 29th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3–10, 2006.

[4] Chiwon Ahn. Exploring chatgpt for information of cardiopulmonary resuscitation. *Resuscitation*, 185, 2023.

[5] Reid Andersen, Christian Borgs, Jennifer Chayes, John Hopcroft, Kamal Jain, Vahab Mirrokni, and Shanghua Teng. Robust pagerank and locally computable spam detection features. In *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pages 69–76, 2008.

[6] Avi Arampatzis, Jaap Kamps, and Stephen Robertson. Where to stop reading a ranked list? Threshold optimization using truncated score distributions. In *Proceedings of the 32nd Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 524–531, 2009.

[7] Avi Arampatzis and Stephen Robertson. Modeling score distributions in information retrieval. *Information Retrieval*, 14:26–46, 2011.

[8] Avi Arampatzis, Stephen Robertson, and Jaap Kamps. Score distributions in information retrieval.

[9] Javed A Aslam and Mark Montague. Models for metasearch. In *Proceedings of the 24th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 276–284, 2001.

[10] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a competition for cleaning web pages. In *Lrec*, Marrakech, 2008.

[11] Brian Bartell, Garrison W Cottrell, and Richard Belew. Learning to retrieve information. In *Proceedings of the Swedish Conference on Connectionism*, page 27, 1995.

[12] Bogdan Batrinca and Philip C Treleaven. Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, 30(1):89–116, 2015.

[13] Daniel Bauer, Judith Degen, Xiaoye Deng, Priska Herger, Jan Gasthaus, Eugenie Giesbrecht, Lina Jansen, Christin Kalina, Thorben Kräger, Robert Märtin, et al. Fiasco: Filtering the internet by automatic subtree classification, osnabruck. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop, incorporating CleanEval*, volume 4, pages 111–121. Presses univ. de Louvain, 2007.

[14] Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-Yates, and Stefano Leonardi. Link analysis for web spam detection. *ACM Transactions on the Web (TWEB)*, 2(1):1–42, 2008.

[15] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[16] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. *Advances in neural information processing systems*, 13, 2000.

[17] Rodger Benham and J Shane Culpepper. Risk-reward trade-offs in rank fusion. In *Proceedings of the 22nd Australasian Document Computing Symposium*, pages 1–8, 2017.

[18] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic chatnoir: Search engine for the clueweb and the common crawl. In *European Conference on Information Retrieval*, pages 820–824. Springer, 2018.

[19] Janek Bevendorff, Michael Völske, Benno Stein, Alexander Bondarenko, Maik Fröbe, Sebastian Günther, and Matthias Hagen. Webis at TREC 2020: Health Misinformation Track. In *Proceedings of the 29th Text REtrieval Conference (TREC)*, 2020.

[20] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. Nudgecred: supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–30, 2021.

[21] Markus Bink, Steven Zimmerman, and David Elsweiler. Featured snippets and their influence on users' credibility judgements. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 113–122, 2022.

[22] Som S Biswas. Potential use of chat gpt in global warming. *Annals of Biomedical Engineering*, pages 1–2, 2023.

[23] Allan Borodin, Gareth O Roberts, Jeffrey S Rosenthal, and Panayiotis Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology (TOIT)*, 5(1):231–297, 2005.

[24] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv:1508.05326*, 2015.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[26] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd Int. Conf. on Machine learning*, pages 89–96, 2005.

[27] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[28] José Ramom Pichel Campos, Pablo Gamallo Otero, and Iñaki Alegria Loinaz. Measuring diachronic language distance using perplexity: Application to english, portuguese, and spanish. *Natural Language Engineering*, 26(4):433–454, 2020.

[29] Yunbo Cao, Jun Xu, Tie-Yan Liu, Hang Li, Yalou Huang, and Hsiao-Wuen Hon. Adapting ranking svm to document retrieval. In *Proceedings of the 29th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 186–193, 2006.

[30] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th Int. Conf. on Machine Learning*, pages 129–136, 2007.

[31] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv:1708.00055*, 2017.

[32] Olivier Chapelle and Yi Chang. Yahoo! learning to rank challenge overview. In *Proceedings of the Learning to Rank Challenge*, pages 1–24. PMLR, 2011.

[33] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[34] Jose M Chenlo, Javier Parapar, David E Losada, and José Santos. Finding a needle in the blogosphere: An information fusion approach for blog distillation search. *Information Fusion*, 23:58–68, 2015.

[35] Mahesh V Chitrao and Ralph Grishman. Statistical parsing of messages. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[36] Noam Chomsky. Syntactic structures. In *Syntactic Structures*. De Gruyter Mouton, 2009.

[37] Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. In *Proceedings of the 22nd Int. Conf. on Machine Learning*, pages 137–144, 2005.

[38] Wei Chu, Zoubin Ghahramani, and Christopher KI Williams. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6(7):1019–1041, 2005.

[39] Wei Chu and S Sathiya Keerthi. New approaches to support vector ordinal regression. In *Proceedings of the 22nd Int. Conf. on Machine Learning*, pages 145–152, 2005.

[40] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[41] Charles Clarke, Maria Maistro, and Mark Smucker. Overview of the TREC 2021 Health Misinformation Track. In *Proceedings of the 30th Text REtrieval Conference (TREC)*, 2021.

[42] Charles Clarke, Maria Maistro, Mark Smucker, and Guido Zuccon. Overview of the TREC 2020 Health Misinformation Track. In *Proceedings of the 29th Text REtrieval Conference (TREC)*, 2020.

[43] Charles LA Clarke, Mark D Smucker, and Alexandra Vtyurina. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 225–234, 2020.

[44] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Charlie Clarke, and Ellen M. Voorhees. Overview of the TREC 2013 web track. In *Proceedings of the 22nd Text REtrieval Conference, (TREC '13)*, Gaithesburg, Maryland, USA, 2013. National Institute of Standards and Technology (NIST).

[45] European Commission. Flash eurobarometer 404: European citizens' digital health literacy, 2014.

[46] BNC Consortium et al. British national corpus. In *Oxford Text Archive Core Collection*, Oxford, UK, 2007. University of Oxford.

[47] W Bruce Croft, Donald Metzler, and Trevor Strohman. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading, 2010.

[48] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[49] J. De Borda. Mémoire sur les élections au scrutin. *Histoire de l'Academie Royale des Sciences pour 1781 (Paris, 1784)*, 1784.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[51] Matthew S Eastin. Credibility assessments of online health information: The effects of source expertise and knowledge of content. *Journal of Computer-Mediated Communication*, 6(4):JCMC643, 2001.

[52] Gunther Eysenbach. Infodemiology: The epidemiology of (mis) information. *The American Journal of Medicine*, 113(9):763–765, 2002.

[53] Absalom E Ezugwu, Abiodun M Ikotun, Olaide O Oyelade, Laith Abualigah, Jeffery O Agushaka, Christopher I Eke, and Andronicus A Akinyelu. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743, 2022.

[54] Marcos Fernández-Pichel, David E Losada, Juan C Pichel, and David Elsweiler. Reliability prediction for health-related content: a replicability study. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II 43*, pages 47–61. Springer, 2021.

[55] Marcos Fernández-Pichel, Rodrigo Martínez-Castaño, David E Losada, and Juan C Pichel. eXtream: a System for Real-time Monitoring of Dynamic Web Sources. In *Proceedings of the Joint Conference of the Information Retrieval Communities in Europe (CIRCLE 2020). http://ceurws. org*, volume 2621, 2020.

[56] Marcos Fernández-Pichel, Selina Meyer, Markus Bink, Alexander Frummet, David E Losada, and David Elsweiler. Improving the reliability of health information credibility assessments. 2023.

[57] Marcos Fernández-Pichel, Manuel Prada-Corral, David E Losada, and Juan C Pichel. CiTIUS at the TREC 2022 Health Misinformation Track. In *Proceedings of the 31st Text REtrieval Conference (TREC)*, 2022.

[58] Marcos Fernández-Pichel, Manuel Prada-Corral, David E Losada, Juan C Pichel, and Pablo Gamallo. CiTIUS at the TREC 2021 Health Misinformation Track. In *Proceedings of the 30th Text REtrieval Conference (TREC)*, 2021.

[59] Aidan Finn, Nicholas Kushmerick, and Barry Smyth. Fact or fiction: Content classification for digital libraries. In *DELOS*, 2001.

[60] Brian J Fogg. Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pages 722–723, 2003.

[61] Brian J Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pages 1–15, 2003.

[62] Forbes. Introducing chatgpt, November 2022. [accessed April 4, 2023].

[63] Forbes. Large language models will define artificial intelligence, January 2023. [accessed March 14, 2023].

[64] Edward A Fox and Joseph A Shaw. Combination of multiple searches. *NIST special publication SP*, 243:243–252, 1994.

[65] Susannah Fox. *Health topics: 80% of internet users look for health information online*. Pew Internet & American Life Project, 2011.

[66] Susannah Fox and Maeve Duggan. Health online 2013. *Health*, 2013:1–55, 2013.

[67] Pablo Gamallo., Manuel Corral., and Marcos Garcia. Comparing dependency-based compositional models with contextualized word embeddings. In *Proceedings of the 13th Int. Conf. on Agents and Artificial Intelligence (ICAART) - Volume 2*, pages 1258–1265. SciTePress, 2021.

[68] Pablo Gamallo, José Ramom Pichel, and Iñaki Alegria. From Language Identification to Language Distance. *Physica A*, 484:162–172, 2017.

[69] Gerald Gazdar, Roger Evans, Alex Franz, and Karen Osborne. *Natural language processing in the 1980s: a bibliography*. CSLI Publications, 1987.

[70] Shalini Ghosh, Oriol Vinyals, Brian Strope, Scott Roy, Tom Dean, and Larry Heck. Contextual lstm (clstm) models for large scale nlp tasks. *arXiv preprint arXiv:1602.06291*, 2016.

[71] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. *Found. Trends Inf. Retr.*, 9(5):355–475, December 2015.

[72] Alexandru L. Ginsca, Adrian Popescu, and Mihai Lupu. Credibility in information retrieval. *Found. Trends Inf. Retr.*, 9(5):355–475, December 2015.

[73] Meritxell González. An analysis of twitter corpora and the differences between formal and colloquial tweets. In *Proceedings of the Tweet Translation Workshop 2015*, pages 1–7, Alicante, Spain, 2015.

[74] Kathleen M Griffiths, Thanh Tin Tang, David Hawking, and Helen Christensen. Automated assessment of the quality of depression websites. *Journal of Medical Internet Research*, 7(5):e59, 2005.

[75] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. Tweetcred: Real-time credibility assessment of content on twitter. In *International conference on social informatics*, pages 228–243. Springer, 2014.

[76] Carolin Hahnel, Frank Goldhammer, Ulf Kröhne, and Johannes Naumann. The role of reading skills in the evaluation of online information gathered from search engine environments. *Computers in Human Behavior*, 78:223–234, 2018.

[77] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[78] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[79] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[80] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[81] T Ryan Hoens and Nitesh V Chawla. Imbalanced datasets: from sampling to classifiers. *Imbalanced learning: Foundations, algorithms, and applications*, pages 43–59, 2013.

[82] Amir Hussain, Erik Cambria, Soujanya Poria, Ahmad Hawalah, and Francisco Herrera. Information fusion for affective computing and sentiment analysis. *Information Fusion*, 71:97–98, 2021.

[83] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, 2014.

[84] Md Saiful Islam, Tonmoy Sarkar, et al. Covid-19–related infodemic and its impact on public health: A global social media analysis. *The American Journal of Tropical Medicine and Hygiene*, 103(4):1621–1629, 2020.

[85] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

[86] Jimmy Jimmy, Guido Zuccon, Joao Palotti, Lorraine Goeuriot, and Liadh Kelly. Overview of the CLEF 2018 Consumer Health Search task. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2018.

[87] Thorsten Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods: support vector learning*, pages 169–184. MIT Press, 1999.

[88] Michał Kąkol, Michał Jankowski-Lorek, Katarzyna Abramczuk, Adam Wierzbicki, and Michele Catasta. On the subjectivity and bias of web content credibility evaluations. In *Proceedings of the 22nd international conference on world wide web*, pages 1131–1136, 2013.

[89] Michal Kakol, Radoslaw Nielek, and Adam Wierzbicki. Understanding and predicting web content credibility using the content credibility corpus. *Information Processing & Management*, 53(5):1043–1061, 2017.

[90] Chris Kamphuis, Arjen P de Vries, Leonid Boytsov, and Jimmy Lin. Which BM25 do you mean? a large-scale reproducibility study of scoring variants. In *European Conference on Information Retrieval*, pages 28–34. Springer, 2020.

[91] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, M Nithya, S Kannan, and V Gurusamy. Preprocessing techniques for text mining. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2014.

[92] Markus Kattenbeck and David Elsweiler. Understanding credibility judgements for web search snippets. *Aslib Journal of Information Management*, 2019.

[93] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 441–450, New York, USA, 2010.

[94] Peter A Lachenbruch. Mcnemar test. *Wiley StatsRef: Statistics Reference Online*, 2014.

[95] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[96] Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*, 2020.

[97] Jurek Leonhardt, Avishek Anand, and Megha Khosla. Boilerplate removal using a neural sequence labeling model. In *Companion Proceedings of the Web Conference 2020*, pages 226–229, Taipei, Taiwan, 2020.

[98] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[99] Q Vera Liao and Wai-Tat Fu. Age differences in credibility judgments of online health information. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(1):1–23, 2014.

[100] Lucas Chaves Lima, Dustin Brandon Wright, Isabelle Augenstein, and Maria Maistro. University of Copenhagen participation in TREC Health Misinformation Track 2020. *arXiv:2103.02462*, 2021.

[101] Christina Lioma, Jakob Grue Simonsen, and Birger Larsen. Evaluation measures for relevance and credibility in ranked lists. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 91–98, 2017.

[102] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021.

[103] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[104] Tie-Yan Liu. *Learning to rank for information retrieval*. Springer Science & Business Media, 2011.

[105] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[106] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[107] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

[108] D. Losada and L. Azzopardi. An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval*, 11:109–138, 2008.

[109] David E Losada, Javier Parapar, and Alvaro Barreiro. A rank fusion approach based on score distributions for prioritizing relevance assessments in information retrieval evaluation. *Information Fusion*, 39:56–71, 2018.

[110] Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *arXiv preprint arXiv:2003.11563*, 2020.

[111] Alejandro G Martín, Alberto Fernández-Isabel, César González-Fernández, Carmen Lancho, Marina Cuesta, and Isaac Martín de Diego. Suspicious news detection through semantic and sentiment measures. *Engineering Applications of Artificial Intelligence*, 101:104230, 2021.

[112] Rodrigo Martínez-Castaño, Juan C Pichel, and Pablo Gamallo. Polypus: a big data self-deployable architecture for microblogging text extraction and real-time sentiment analysis. *arXiv preprint arXiv:1801.03710*, 2018.

[113] David Matsumoto, Hyisung C Hwang, and Vincent A Sandoval. Cross-language applicability of linguistic features associated with veracity and deception. *Journal of Police and Criminal Psychology*, 30(4):229–241, 2015.

[114] Scott C Matthews, Alvaro Camacho, Paul J Mills, and Joel E Dimsdale. The internet for medical information about cancer: help or hindrance? *Psychosomatics*, 44(2):100–103, 2003.

[115] Michael McCandless, Erik Hatcher, Otis Gospodnetić, and O Gospodnetić. *Lucene in action*, volume 2. Manning Greenwich, 2010.

[116] D Harrison McKnight and Charles J Kacmar. Factors and effects of information credibility. In *Proceedings of the ninth international conference on Electronic commerce*, pages 423–432, 2007.

[117] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[118] Alistair Moffat and Justin Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)*, 27(1):1–27, 2008.

[119] Alberto Moro and Laura Lonza. Electricity carbon intensity in european member states: Impacts on ghg emissions of electric vehicles. *Transportation Research Part D: Transport and Environment*, 64:5–14, 2018.

[120] Subhabrata Mukherjee and Gerhard Weikum. Leveraging joint interactions for credibility analysis in news communities. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 353–362, 2015.

[121] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*, 2016.

[122] Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. *arXiv:2003.06713*, 2020.

[123] OpenAI. Gpt-4 technical report. *arXiv:submit/4812508*, 2023.

[124] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In *European Conference on Information Retrieval*, pages 517–519. Springer, 2005.

[125] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[126] Javier Parapar, David E Losada, and Álvaro Barreiro. Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, pages 655–664, Virtual Event Republic of Korea, 2021.

[127] Javier Parapar, David E Losada, Manuel A Presedo-Quindimil, and Alvaro Barreiro. Using score distributions to compare statistical significance tests for information retrieval evaluation. *Journal of the Association for Information Science and Technology*, 71(1):98–113, 2020.

[128] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[129] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, 31(7):770–780, 2020.

[130] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[131] Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR Int. Conf. on Theory of Information Retrieval*, pages 209–216, 2017.

[132] Jan Pomikálek. Removing boilerplate and duplicate content from web corpora. *Disertacnı práce, Masarykova univerzita, Fakulta informatiky*, 2011.

[133] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *21st Annual Interaction ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–56. ACM New York, NY, USA, 1998.

[134] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2173–2178, 2016.

[135] Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2066–2070, 2021.

[136] Ronak Pradeep, Xueguang Ma, Xinyu Zhang, Hang Cui, Ruizhou Xu, Rodrigo Nogueira, and Jimmy Lin. H2oloo at TREC 2020: When all you got is a hammer... deep learning, health misinformation, and precision medicine. In *Proceedings of the 29th Text REtrieval Conference (TREC)*, 2020.

[137] Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv:2101.05667*, 2021.

[138] Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, Xu-Dong Zhang, and Hang Li. Learning to search web pages with query-level loss functions. *Technical Report*, 156:28, 2006.

[139] Martin Radfar, Athanasios Mouchtaris, and Siegfried Kunzmann. End-to-end neural transformer based spoken language understanding. *arXiv preprint arXiv:2008.10984*, 2020.

[140] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[141] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[142] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv:1910.10683*, 2019.

[143] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[144] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese bert-networks. *arXiv:1908.10084*, 2019.

[145] Reuters Insitute, University of Oxford. *Reuters Digital News Report*, 2021. [accessed June 9, 2022].

[146] Soo Young Rieh. Judgment of information quality and cognitive authority in the web. *Journal of the American society for Information Science and Technology*, 53(2):145–161, 2002.

[147] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[148] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. Okapi at TREC-3. *NIST Special Publication Sp*, 109:109, 1995.

[149] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[150] Julia Schwarz and Meredith Morris. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1245–1254, 2011.

[151] Rico Sennrich. Perplexity minimization for translation model domain adaptation in statistical machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[152] Karishma Sharma, Sungyong Seo, Chuizheng Meng, Sirisha Rambhatla, and Yan Liu. Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020.

[153] Rosa Sicilia, Mario Merone, Roberto Valenti, and Paolo Soda. Rule-based space characterization for rumour detection in health. *Engineering Applications of Artificial Intelligence*, 105:104389, 2021.

[154] Thamar Solorio, Melissa Sherman, Yang Liu, Lisa M Bedore, Elisabeth D Peña, and Aquiles Iglesias. Analyzing language samples of spanish–english bilingual children for the automated prediction of language dominance. *Natural Language Engineering*, 17(3):367–395, 2011.

[155] Parikshit Sondhi, VG Vinod Vydiswaran, and ChengXiang Zhai. Reliability prediction of webpages in the medical domain. In *European Conference on Information Retrieval*, pages 219–231. Springer, 2012.

[156] Miroslav Spousta, Michal Marek, and Pavel Pecina. Victor: the web-page cleaning tool. In *4th Web as Corpus Workshop (WAC4)-Can we beat Google*, pages 12–17, 2008.

[157] Aixin Sun, Ee-Peng Lim, and Wee-Keong Ng. Web classification using support vector machine. In *Proceedings of the 4th international workshop on Web information and data management*, pages 96–99, McLean Virginia, USA, 2002.

[158] Nigar M Shafiq Surameery and Mohammed Y Shakor. Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290*, 3(01):17–22, 2023.

[159] Julian Unkel and Alexander Haas. The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology*, 68(8):1850–1862, 2017.

[160] Elizabeth B Varghese and Sabu M Thampi. A multimodal deep fusion graph framework to detect social distancing violations and fcgs in pandemic surveillance. *Engineering Applications of Artificial Intelligence*, 103:104305, 2021.

[161] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[162] Neil Vigdor. Man fatally poisons himself while self-medicating for coronavirus, doctor says, March 2020. [Online; posted 24-March-2020].

[163] Marco Viviani and Gabriella Pasi. Credibility in social media: opinions, news, and health information—a survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 7(5):e1209, 2017.

[164] Thijs Vogels, Octavian-Eugen Ganea, and Carsten Eickhoff. Web2text: Deep structured boilerplate removal. In *European Conference on Information Retrieval*, pages 167–179, Grenoble, France, 2018. Springer.

[165] VG Vinod Vydiswaran, ChengXiang Zhai, and Dan Roth. Content-driven trust propagation framework. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 974–982, 2011.

[166] Haifeng Wang, Lvjiyuan Jiang, Qian Zhao, Hao Li, Kai Yan, Yang Yang, Songlin Li, Yungang Zhang, Lianliu Qiao, Cuilian Fu, et al. Progressive structure network-based multiscale feature fusion for object detection in real-time application. *Engineering Applications of Artificial Intelligence*, 106:104486, 2021.

[167] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.

[168] Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F Wong, and Lidia S Chao. Learning deep transformer models for machine translation. *arXiv preprint arXiv:1906.01787*, 2019.

[169] C Nadine Wathen and Jacquelyn Burkell. Believe it or not: Factors influencing credibility on the web. *Journal of the American society for information science and technology*, 53(2):134–144, 2002.

[170] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[171] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

[172] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[173] Chuan Wu, Evangelos Kanoulas, and Maarten de Rijke. It all starts with entities: A salient entity topic model. *Natural Language Engineering*, 26(5):531–549, 2020.

[174] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1253–1256, New York, NY, USA, 2017. Association for Computing Machinery.

[175] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Enabling the use of lucene for information retrieval research. In *Proceedings of the 40th Int. ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1253–1256, 2017.

[176] Peilin Yang, Hui Fang, and Jimmy Lin. Anserini: Reproducible ranking baselines using lucene. *Journal of Data and Information Quality (JDIQ)*, 10(4):1–20, 2018.

[177] Wei Zha and H Denis Wu. The impact of online disruptive ads on users' comprehension, evaluation of site credibility, and sentiment of intrusiveness. *American Communication Journal*, 16(2), 2014.

[178] Yu Zhang, Houquan Zhou, and Zhenghua Li. Fast and accurate neural crf constituency parsing. *arXiv:2008.03736*, 2020.

[179] Guido Zuccon and Bevan Koopman. Dr chatgpt, tell me what i want to hear: How prompt knowledge impacts health answer correctness. *arXiv preprint arXiv:2302.13793*, 2023.

List of Figures

Fig. 1.1 Licensing information of the previous Elsevier publication. 11
Fig. 1.2 The previous paper has been published as open access. 11
Fig. 1.3 Licensing information of a Springer publication. 12
Fig. 1.4 Licensing information of an ACM publication. 13
Fig. 1.5 Licensing information of a CEUR-WS publication (CIRCLE 2022). 15
Fig. 1.6 Licensing information of a CEUR-WS publication (CIRCLE 2020). 15
Fig. 1.7 Licensing information of a CEUR-WS publication (ROMCIR 2023). 16

Fig. 2.1 Document-term matrix standardisation. 25
Fig. 2.2 A TREC 2020 Health Misinformation Track topic (Topic 13). 29
Fig. 2.3 Variation of the F1 macro precision with percent training data used in trustworthiness and readability tasks. 39
Fig. 2.4 Variation of the training time (ms) with percent training data used in trustworthiness and readability tasks. *Y* axis in log scale. 39

Fig. 3.1 Full pipeline for health-related misinformation detection. After indexing the corpus, the system supports a document retrieval stage, passage-based re-ranking of the top retrieved documents, passage reliability estimation, and a final re-ranking stage that combines multiple signals. 44
Fig. 3.2 Document retrieval phase. The corpus is indexed with Anserini and, next, queries are executed against the resulting indexing. Search is done with the BM25 implementation from Pyserini. 44

Fig. 3.3 T5 fine-tuning for ranking passages (example in the upper part from Raffel et al.[142]). The pre-training stage tunes the model for general language understanding tasks and, next, the model is fine-tuned for the estimation of relevance at passage level. 46
Fig. 3.4 Passage re-ranking phase. The top retrieved documents from the initial ranking (R_0) are re-ordered based on the most relevant passages (passage relevance estimates obtained from MonoT5). 47
Fig. 3.5 Sliding window ($window = 6$ and $stride = 3$ sentences) used to perform passage re-ranking. 47
Fig. 3.6 T5 fine-tuning process for passage reliability classification. The fine-tuning stage takes queries and passages labelled in terms of reliability. 48
Fig. 3.7 Unsupervised strategy for passage reliability detection. Sentences from the most relevant passages are represented with Sentence BERT and their similarity to the Sentence BERT representation of the query expression is computed. 48
Fig. 3.8 Logistic regression weights obtained from each training fold. 63
Fig. 3.9 ad-hoc results: Compatibility of runs with helpful and harmful results. A good run is helpful and not harmful. For a certain level of helpfulness, less harm is prefered. 65

Fig. 5.1 Social Minder architecture. 80
Fig. 5.2 Social Minder dashboard (upper part). 81
Fig. 5.3 Social Minder dashboard (bottom part). 81
Fig. 5.4 Perplexity model for boilerplate removal. 85
Fig. 5.5 Space savings derived from using our model in the IR search task. 92
Fig. 5.6 Some of the Pyplexity capabilities shown up and running in a Jupyter Notebook. 97
Fig. 5.7 Demo webpage that showcases the capabilities of the cleaning tool presented in this section. 98

List of Tables

Tab. 2.1 Class distribution in Sondhi's dataset. 20
Tab. 2.2 Sondhi et al. original paper results. 23
Tab. 2.3 Our results for Sondhi et al. dataset. 24
Tab. 2.4 Our results for Sondhi et al. dataset (with standard scaler). 25
Tab. 2.5 Class distribution in the different datasets. 26
Tab. 2.6 Our results for Schwarz et al. dataset. 27
Tab. 2.7 Our results for CLEF eHealth dataset. 27
Tab. 2.8 Our results for the Total Recall Task. 31
Tab. 2.9 Our results for the AdHoc Retrieval Task. 31
Tab. 2.10 Label distribution in the CLEF eHealth dataset. 33
Tab. 2.11 Trustworthiness results obtained when setting or not the cost-factor to the proportion between classes. 36
Tab. 2.12 Readability results obtained when setting or not the cost-factor to the proportion between classes. 37
Tab. 2.13 Usefulness results obtained when setting or not the cost-factor to the proportion between classes. 38

Tab. 3.1 Preference ordering for documents mapped to graded relevance. Usefulness=1 (0) means that the document is on-topic (off-topic). Correctness=1 (0) means that the document gives a correct (incorrect) answer to the health-related request. Correctness=2 means that the document gives no answer to the health-related request. Correctness=-1 means that the document was not manually judged in terms of correctness to the health-related request. Credibility=1 (0) means that the document was judged as credible (non-credible). Credibility=-1 means that the document was not judged in terms of credibility. 54
Tab. 3.2 Relevance-based search method results for the total recall task. 56
Tab. 3.3 Relevance-based search method results for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value. 56
Tab. 3.4 Passage reliability estimation results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not. 58
Tab. 3.5 Passage reliability estimation results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value. 58
Tab. 3.6 CombSUM results (supervised and unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not. 59
Tab. 3.7 CombSUM results (supervised and unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value. . . 60
Tab. 3.8 Borda Count and Learning-to-Rank results (only unsupervised methods) for the total recall task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not. 61
Tab. 3.9 Borda and Learning-to-Rank results (only unsupervised methods) for the ad-hoc retrieval task. Please note that the ↑/↓ symbols indicate whether the method significantly improves the baseline or not, but in the case of the Compatibility harmful measure, an improvement is associated to a lower absolute value. . . 61
Tab. 3.10 Comparison of official TREC 2020 runs and our best performers for the total recall task (data extracted from [42]). 64

Tab. 3.11 Comparison of official TREC 2020 runs and our best performers for the ad-hoc retrieval task (data extracted from [42]). For the sake of simplicity we only report here the official metric by which the participating solutions were ranked (the difference between compatibility values). 64

Tab. 4.1 Zero-shot experiments, proportion of correct answers of each model-prompt combination for the three TREC datasets. The best performing prompt for each model is marked in bold. The last two rows report the average performance and the standard deviation of each model across all prompts. 73

Tab. 4.2 Few-shot experiments, proportion of correct answers of each model-prompt combination with varying number of in-context examples. For each row, the best score is marked in bold and the symbol * marks those cases where McNemar's test ($\alpha = .05$) finds a difference between the corresponding model and its 0-shot equivalent. 75

Tab. 5.1 General statistics of search collections. 86

Tab. 5.2 TREC 2019 Decision Track. Effect of different perplexity cutoff values (reported in brackets) on retrieval performance. The ↑/↓ symbols indicate whether the method significantly improves or not (Wilcoxon test, $\alpha = 0.05$) over the baseline (no perplexity-based removal). 90

Tab. 5.3 TREC 2013 Web Track. The ↑/↓ symbols indicate whether the method significantly improves or not (Wilcoxon test, $\alpha = 0.05$) over the baseline (no perplexity-based removal). 92

Tab. 5.4 Time measurements (HH:MM:SS) for the baseline and some perplexity-based variants. 93

Tab. 5.5 Carbon emissions in kg of CO_2 for the baseline and some perplexity-based variants. 94

Tab. 5.6 Classification results for WebKB dataset. For each block, the top performer (F1 macro) is bolded. 94

Tab. 5.7 CleanEval shared-task results for two perplexity-based methods and the jusText algorithm. The percentage scores represent the average similarity between the webpages cleaned automatically and the ground truth webpages. . . 95

The evolution of the Web has led to an improvement in information accessibility. This change has allowed access to more varied content at greater speed, but we must also be aware of the dangers involved. The results offered may be unreliable, inadequate, or of poor quality, leading to misinformation. This can have a greater or lesser impact depending on the domain, but is particularly sensitive when it comes to health-related content.

In this thesis, we focus in the development of methods to automatically assess credibility. We also studied the reliability of the new Large Language Models (LLMs) to answer health questions. Finally, we also present a set of tools that might help in the massive analysis of web textual content.