UNIVERSITY OF MILANO-BICOCCA
Department of Informatics, Systems and Communication
Master's degree program in Data Science

Agenzia Nazionale Stampa Associata

Multi-domain claim detection:

A coreset and an external feature based approach for automated fact-checking

Supervisor: Prof. Gabriella Pasi
Co-supervisor: Prof. Marco Viviani
Co-supervisor: Dr. Sandip Modha
Co-supervisor: Dr. Marco Beltrame

Master's degree thesis by **Tommaso Redaelli**

Academic Year 2023-2024

Motivations [1] [2] [3]

Risks

Distortion of public opinion

Misleading information can shape people's views incorrectly.

Political manipulation

Fake news can skew political debates and electoral outcomes.

Public safety threats

False information may lead to harmful actions or widespread panic.

Usefulness

Preserving information integrity

Ensures accurate, reliable data in the public domain.

Enhancing public discourse

Supports healthy, fact-based discussions and debates.

Strengthening democracy

Protects democratic processes by promoting informed decision-making.

Automated fact checking [3] [4] [5] [6] [7] [8]

01. Claim detection

Identifying factual, verifiable claims whose veracity is of interest or harmful for the public opinion.

02. Evidence retrieval

Retrieval of certified information that can be useful for verifying what is stated in the claims

03. Fact verification

Verify the veracity of the claim based on the evidence collected in the previous step

04. Justification

Producing a coherent and evidence supported justification for the verdict choice

Claim detection [9] [10] [11]

“ The process of selecting claims for verification ”

Main concepts

01. Claim definition

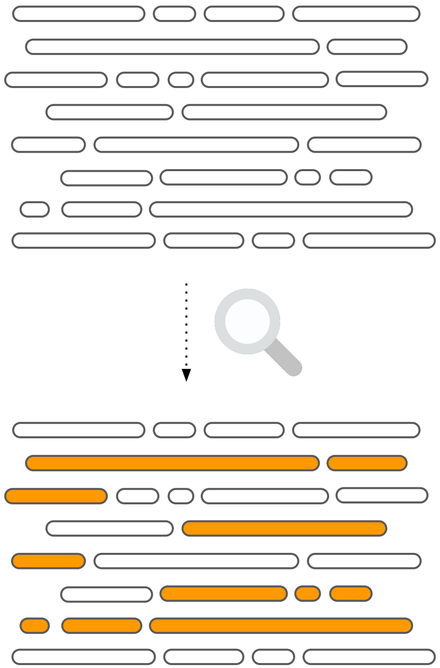
A factual statement that can be verified as true or false.

02. Objectivity

Impartiality in the content of the statement, no subjective opinions.

03. Check worthiness

Public opinion is interested in knowing the veracity



Main issues

01. Linguistic complexity

Complex language structures, such as metaphors, sarcasm, and irony.

02. Explicit vs. Implicit

Claims can be indirectly tied to verifiable facts.

03. Lexical and context diversity

Different lexical form and structures across different contexts and sources

Objectives and contributions

01.

Generalization

Across different linguistic and context domains

02.

Additional features

Combining basic LLM structure with features related to claim detection issues

03.

Comparison

Across datasets representativeness and different models' performances

Project pipeline

Datasets selection

Generalizing across different linguistic styles and thematic domanis

1.

LLMs comparison

Fine-tuning and evaluating different pre-trained LLM

2.

Empirical evaluation of representativeness

Cross-dataset evaluation revealing the compatibility and diversity between different data sources and models.

3.

Coreset selection

Extracting reduced but representative subsets from datasets

4.

Increase complexity

Introducing additional features

5.

Syntactic features

Exploiting syntactic relations between terms

6.

Concepts’ definitions

Providing definitions of the key concepts in a text

Tackling generalization — *Datasets selection*

[12] [13] [14] [15] [16]

CLEF *CheckThat!* Lab

Dataset from Task 1 in 2020, 2021, 2022, 2023 and 2024

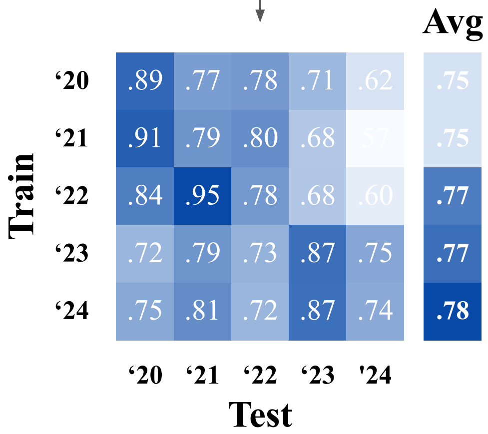
Dataset	Sources	Topics
CheckThat! 2020	*Twitter*	*U.S. election*
CheckThat! 2021	*Speech transcription*	*Politics*
CheckThat! 2022	*Twitter*	*Covid-19*
CheckThat! 2023	*Newspapers*	*Politics*
CheckThat! 2024	*Newspapers, speech transcription, online forums*	*Politics, global emergencies*

LLM selection — *First comparative evaluation*

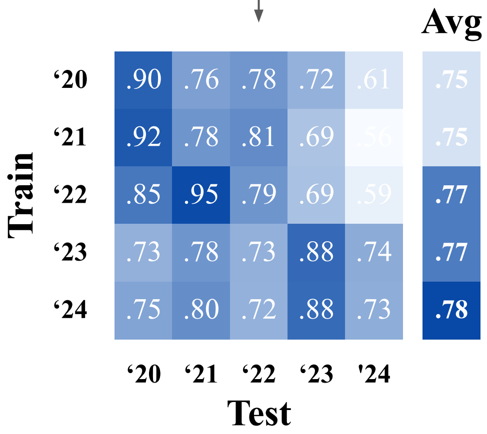
[17] [18] [19] [20] [21]
[22] [23] [24] [25]

Model	Avg. F1
BERT 110 M params – 3.5 h tr. time	**0.819**
XLM-RoBERTa 278 M params – 4.5 h tr. time	0.803
mBERT 177 M params – 4.0 h tr. time	0.780
BART 610 M params – 5.0 h tr. time	0.765
GPT-2 108 M params – 3.5 h tr. time	**0.822**

BERT
cross dataset
evaluation



GPT-2
cross dataset
evaluation



Coreset extraction — *A smaller but still representative subset* [26]

1. Initial dataset D

$$|D| = N$$

2. Average datapoint

$$\bar{x} = \frac{\sum_{i=0}^{N} x_i}{N}$$

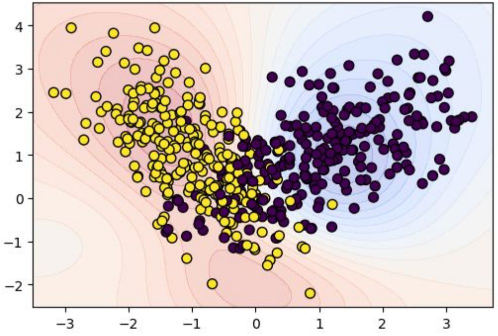
3. Extraction probabilities

$$p_i = x_i - \bar{x}$$

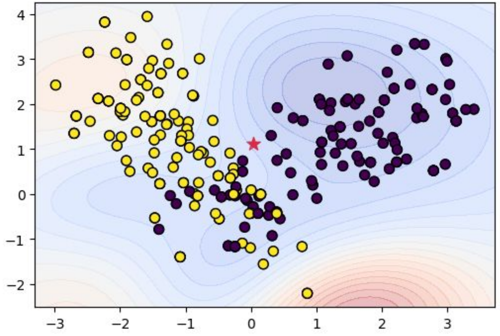
$$p_i = \frac{p_i}{\sum_{j=0}^{N} p_j}$$

4. Coreset C selection

$$|C| = \lfloor 0.5 \cdot N \rfloor$$



Original dataset



Selected coreset

	CT 2020	CT 2021	CT 2022	CT 2023	CT 2024	Avg. F1
BERT + Whole dataset	0.830	**0.943**	0.750	0.860	0.712	0.819
BERT + Whole coreset	**0.848**	0.880	0.772	0.848	0.691	0.808
BERT + Union coreset	0.840	0.884	**0.773**	**0.864**	**0.746**	**0.822**

	CT 2020	CT 2021	CT 2022	CT 2023	CT 2024	Avg. F1
GPT-2 + Whole dataset	0.885	**0.932**	0.763	**0.854**	0.701	0.827
GPT-2 + Whole coreset	0.911	0.873	0.792	0.845	0.682	0.821
GPT-2 + Union coreset	**0.928**	0.868	**0.808**	0.853	**0.710**	**0.833**

Increase complexity — *Exploit additional features* [10] [27] [28] [29]

Rethinking to the specific claim detection task

What can be useful to recognize claims ?

" A claim is an objective, verifiable and free of personal judgments assertion "

Integrate syntactic features

Relevance in **different linguistic structures**

Clarifying **ambiguities**

Support in the analysis of **long texts**

" A claim is an assertion to which public opinion is interested in its veracity "

Provide concepts definitions

Interpretation in relation to the **context**

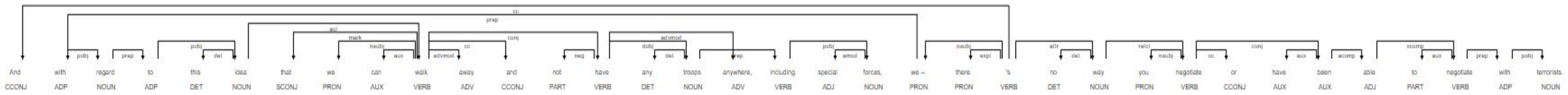
Understanding **complex concepts**

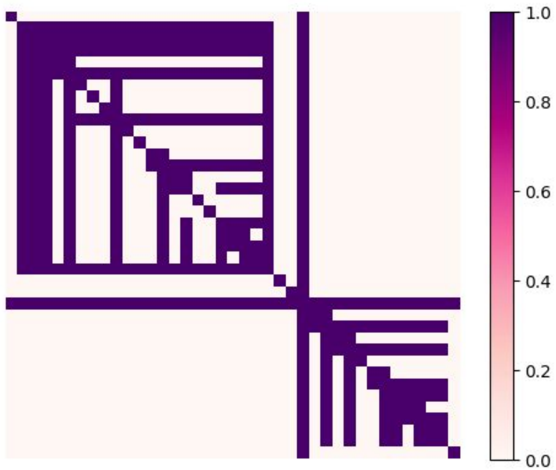
Support in the analysis of **short texts**

Syntactic dependencies — *Representation* [30]

Example text:

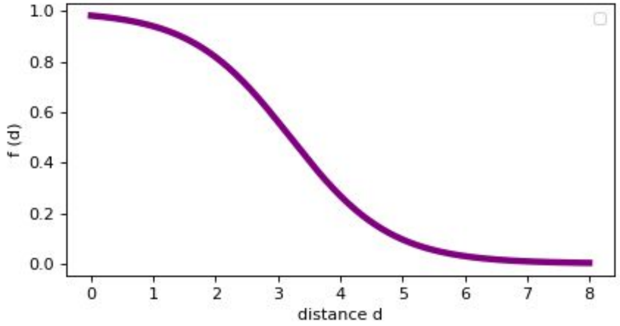
" And with regard to this idea that we can walk away and not have any troops anywhere, including special forces, we -- there's no way you negotiate or have been able to negotiate with terrorists. "

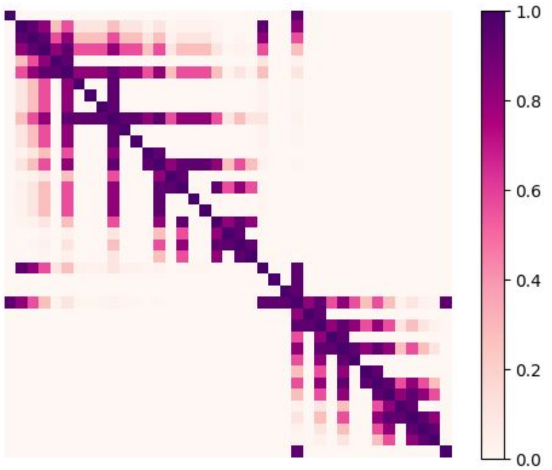




Binary adjacency matrix

$$f(d) = \frac{1}{1 + e^{1.25\,(d-4)}}$$

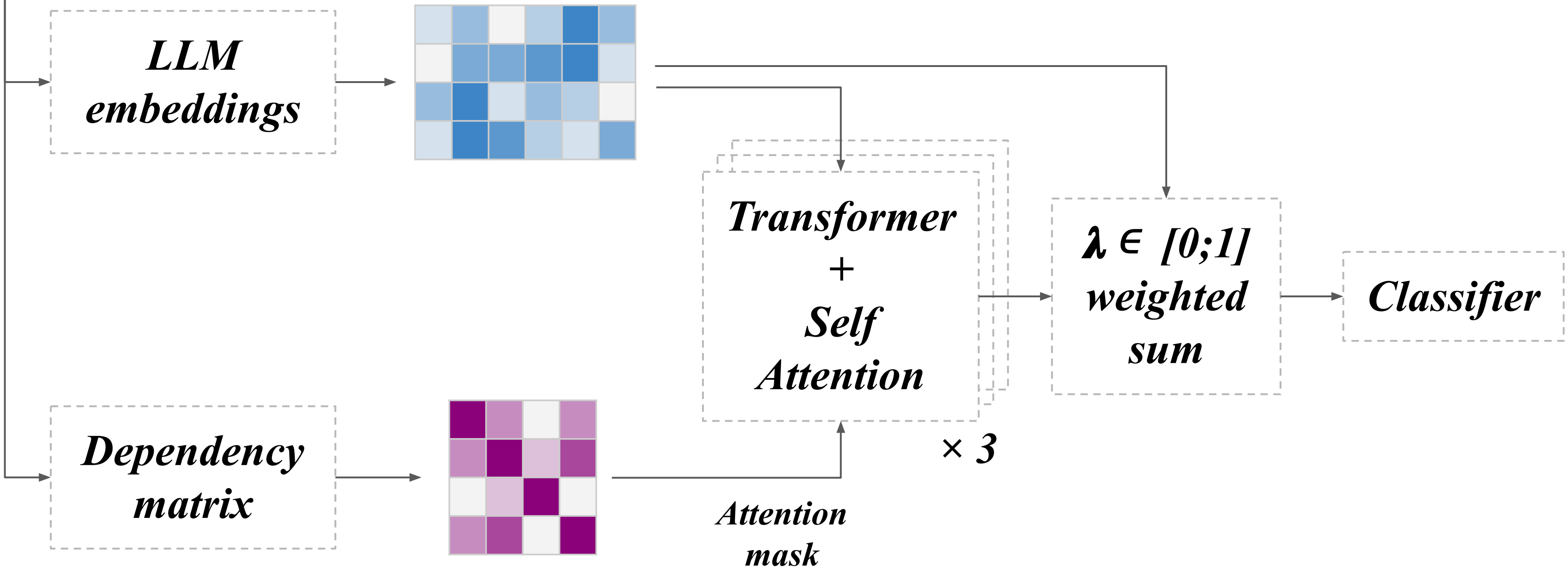




Adjacency matrix with sigmoid correction

Syntactic dependencies — *LLM integration* [31] [32]

Example text:
" If we just wore these masks - the President's own advisors have told him - we could save 100,000 lives. "



Concepts definition — *Extracting text key concepts* [33]

Example text:
"Antifa is an idea not an organization..."

Search PosTag pattern

- *(NOUN | PROPN) +*
- *(ADJ) . (NOUN | PROPN)*

- *" Antifa "*
- *" idea "*
- *" organization "*

Semantic similarity

Top-K (K=2) over
cosine-similarity (BERT[text] ; BERT[term])
∀ term in candidates

- *" Antifa "*
- *" idea "*

Gathering definitions

*Web Scraping on **Wikipedia** or **Google-Search***

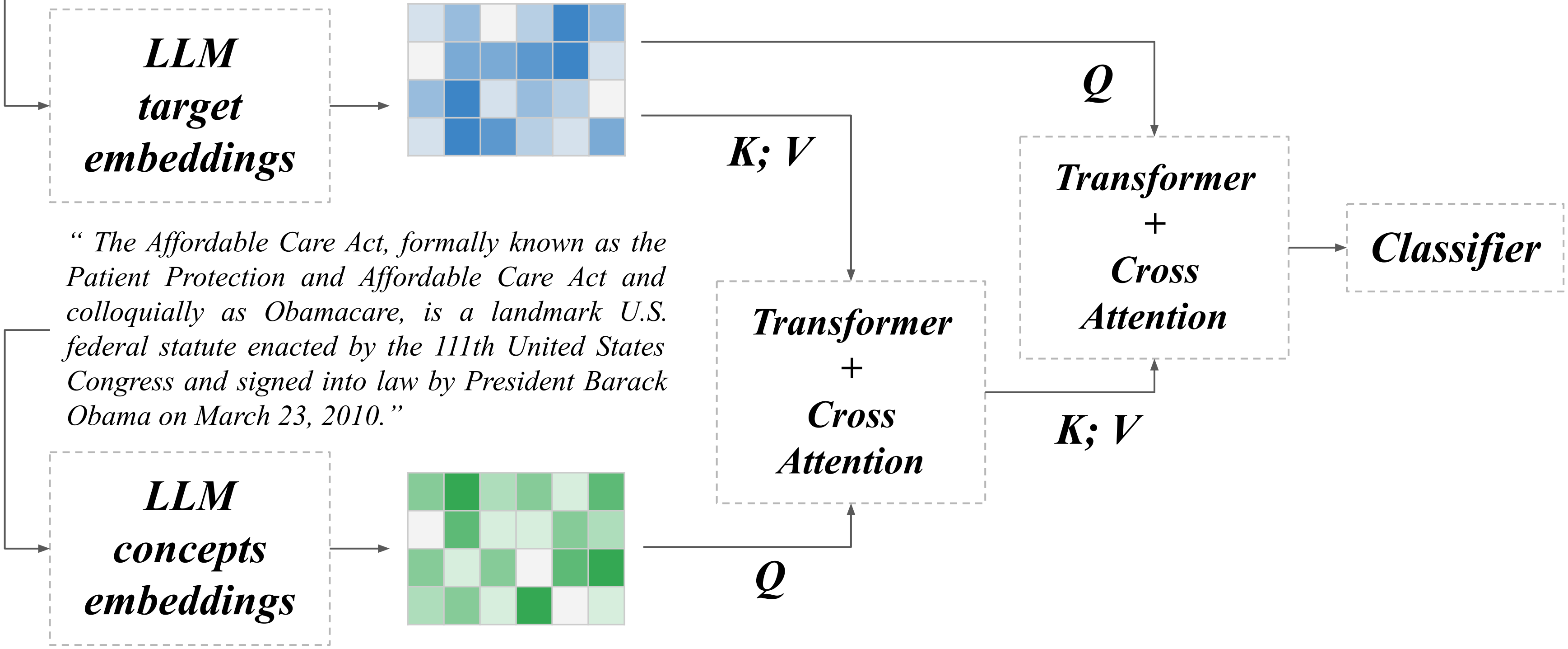
" Antifa " *: " Antifa is a left-wing anti-fascist and anti-racist political movement in the United States. It consists of a highly decentralized array of autonomous groups that use nonviolent direct action [...] to achieve their aims. "*

" idea " *: " a thought or suggestion as to a possible course of action."*

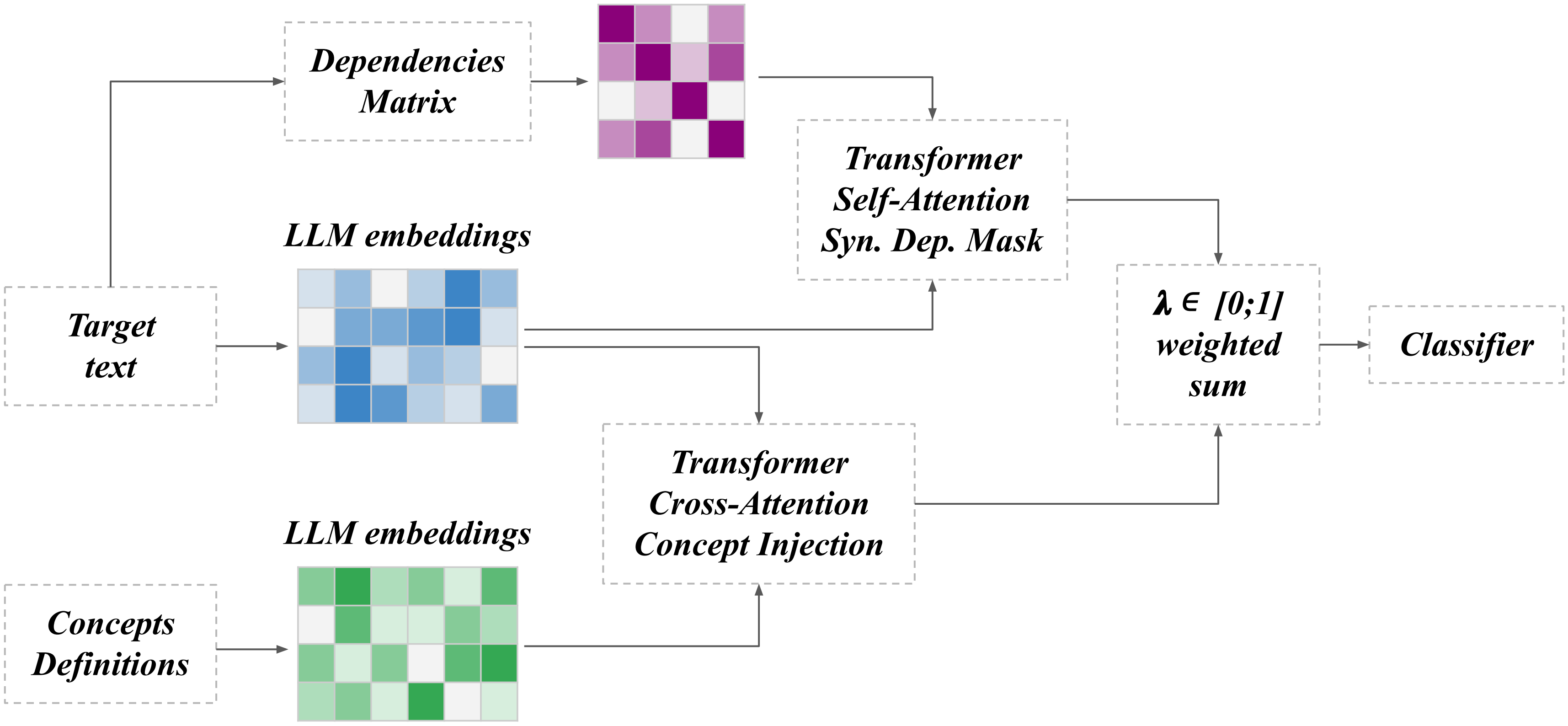
Concepts definition — *LLM injection mechanism* [29]

Example text:

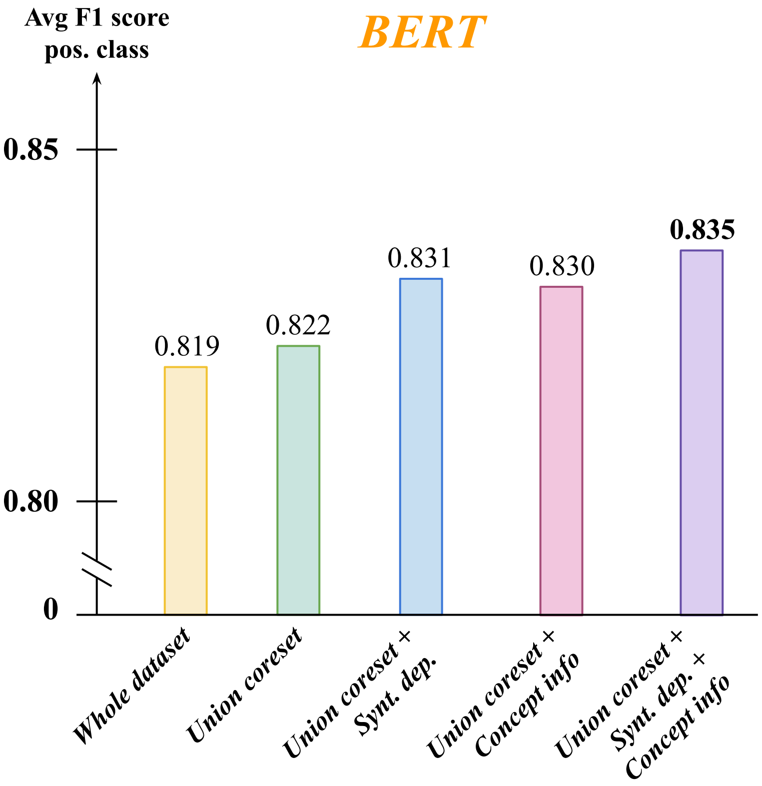
“ The individual mandate was the most unpopular aspect of ***Obamacare****.”*

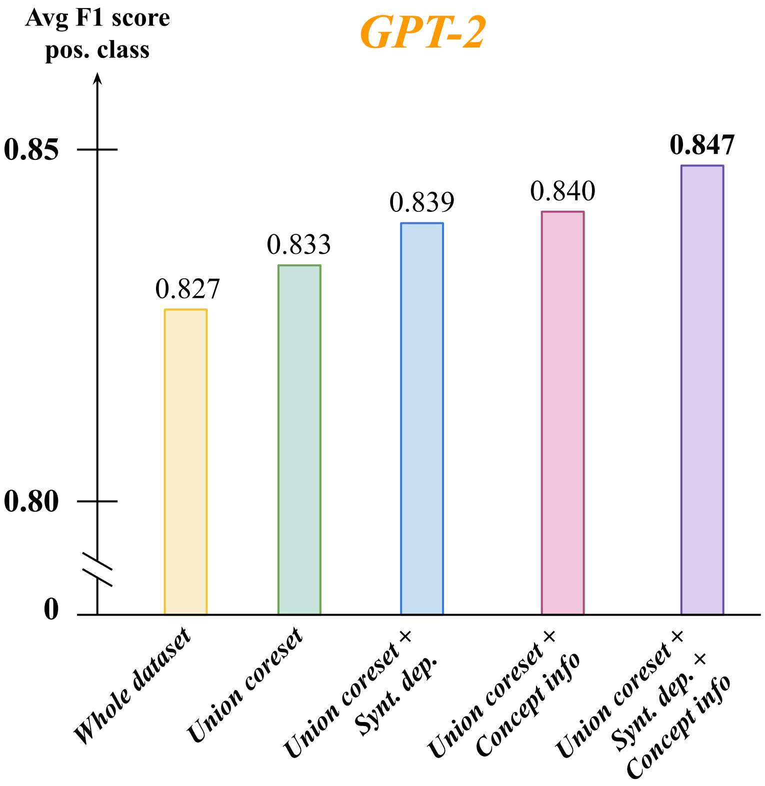


Complete model — *Both syntactic and concepts informations*



Results — *Ablation study on model's components*





Future improvements — *Limitations and possible solutions*

Limits		*Possible solutions*
Coreset selection	Naive method	**Importance sampling** Based on the impact on the cost function
	Fixed size	**Variable dimension** Based on the degree of representation
Syntactic dependencies	Unused logic functions	**Using graph-based neural networks** Edges as categorical features
Concept extraction	Fixed number	**Variable number** Make use of a threshold
	No disambiguation	**Extraction of different definition** Selected by the context of the term within the sentence
		Explore structured knowledge bases Graph-based knowledge bases
Generalization	Single language	**Different databases and models** Fine tune and compare more complex multilingual models

Bibliography — *References to used and related works — 1 / 3*

1. David MJ Lazer et al. "The science of fake news". In: Science 359.6380 (2018), pp. 1094–1096.
2. Xinyi Zhou and Reza Zafarani. "A survey of fake news: Fundamental theories, detection methods, and opportunities". In: ACM Computing Surveys (CSUR) 53.5 (2020), pp. 1–40.
3. Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. "A survey on automated fact-checking". In: Transactions of the Association for Computational Linguistics 10 (2022), pp. 178–206.
4. Naeemul Hassan, Chengkai Li, and Mark Tremayne. "Detecting check-worthy factual claims in presidential debates". In: Proceedings of the 24th acm international on conference on information and knowledge management. 2015, pp. 1835–1838.
5. Pepa Atanasova et al. "Overview of the CLEF-2018 CheckThat! lab on automatic identification and verification of political claims. Task 1: Check-worthiness". In: arXiv preprint arXiv:1808.05542 (2018).
6. William Ferreira and Andreas Vlachos. "Emergent: a novel data-set for stance classification". In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies. ACL. 2016.
7. James Thorne et al. "FEVER: a large-scale dataset for fact extraction and VERification". In: arXiv preprint arXiv:1803.05355 (2018).
8. Kashyap Popat et al. "Declare: Debunking fake news and false claims using evidence-aware deep learning". In: arXiv preprint arXiv:1809.06416 (2018).
9. Naeemul Hassan et al. "Claimbuster: The first-ever end-to-end fact-checking system". In: Proceedings of the VLDB Endowment 10.12 (2017), pp. 1945–1948.
10. Lev Konstantinovskiy et al. "Toward automated factchecking: Developing an annotation schema and benchmark for consistent automated claim detection". In: Digital threats: research and practice 2.2 (2021), pp. 1–16.
11. Chaoyuan Zuo, Ayla Karakas, and Ritwik Banerjee. "A hybrid recognition system for check-worthy claims using heuristics and supervised learning".
12. CLEF2020-CheckThat! - Tasks 1 & 5: Check-Worthiness --- sites.google.com. https://sites.google.com/view/clef2020-checkthat/tasks/tasks-1-5-check-worthiness, [Accessed 08-10-2024]

Bibliography *— References to used and related works — 2 / 3*

13. CLEF2021-CheckThat! - Task 1: Check-Worthiness Estimation --- sites.google.com. https://sites.google.com/view/clef2021-checkthat/tasks/task-1-check-worthiness-estimation, [Accessed 08-10-2024]

14. CLEF2022-CheckThat! - Task 1: Identifying Relevant Claims in Tweets --- sites.google.com. https://sites.google.com/view/clef2022-checkthat/tasks/task-1-identifying-relevant-claims-in-tweets, [Accessed 08-10-2024]

15. CheckThat! --- checkthat.gitlab.io. https://checkthat.gitlab.io/clef2023/task1/, [Accessed 08-10-2024]

16. CheckThat! --- checkthat.gitlab.io. https://checkthat.gitlab.io/clef2024/task1/, [Accessed 08-10-2024]

17. Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: arXiv preprint arXiv:1810.04805 (2018).

18. google-bert/bert-base-uncased · Hugging Face — huggingface.co. https://huggingface.co/google-bert/bert-base-uncased. [Accessed 22-08-2024].

19. Alexis Conneau et al. "Unsupervised cross-lingual representation learning at scale". In: arXiv preprint arXiv:1911.02116 (2019).

20. FacebookAI/xlm-roberta-base · Hugging Face — huggingface.co. https://huggingface.co/FacebookAI/xlm-roberta-base. [Accessed 22-08-2024].

21. google-bert/bert-base-multilingual-cased · Hugging Face — huggingface.co. https://huggingface.co/google-bert/bert-base-multilingualcased. [Accessed 22-08-2024].

22. Mike Lewis et al. "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension". In: arXiv preprint arXiv:1910.13461 (2019).

23. facebook/bart-base · Hugging Face — huggingface.co. https://huggingface.co/facebook/bart-base. [Accessed 22-08-2024].

24. Alec Radford et al. "Language models are unsupervised multitask learners". In: OpenAI blog 1.8 (2019), p. 9.

Bibliography — *References to used and related works — 3 / 3*

25. openai-community/gpt2 · Hugging Face — huggingface.co. https://huggingface.co/openai-community/gpt2. [Accessed 22-08-2024].

26. Olivier Bachem, Mario Lucic, and Andreas Krause. “Practical coreset constructions for machine learning”. In: arXiv preprint arXiv:1703.06476 (2017).

27. Lan You et al. “DRGAT: Dual-relational graph attention networks for aspect-based sentiment classification”. In: Information Sciences 668 (2024), p. 120531.

28. S Supraja and Andy WH Khong. “Quad-Faceted Feature-Based Graph Network for Domain-Agnostic Text Classification to Enhance Learning Effectiveness”. In: IEEE Transactions on Computational Social Systems (2024).

29. Tilman Beck, Andreas Waldis, and Iryna Gurevych. “Robust integration of contextual information for cross-target stance detection”.

30. Xing Zhang et al. “Detecting dependency-related sentiment features for aspect-level sentiment classification”.

31. A Vaswani. “Attention is all you need”. In: Advances in Neural Information Processing Systems (2017).

32. Zhongli Li et al. “Improving BERT with syntax-aware local attention”. In: arXiv preprint arXiv:2012.15150 (2020).

33. Tim Schopf, Simon Klimek, and Florian Matthes. “Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction”. In: arXiv preprint arXiv:2210.05245 (2022).