A Neighbourhood Framework for Resource-Lean Content Flagging

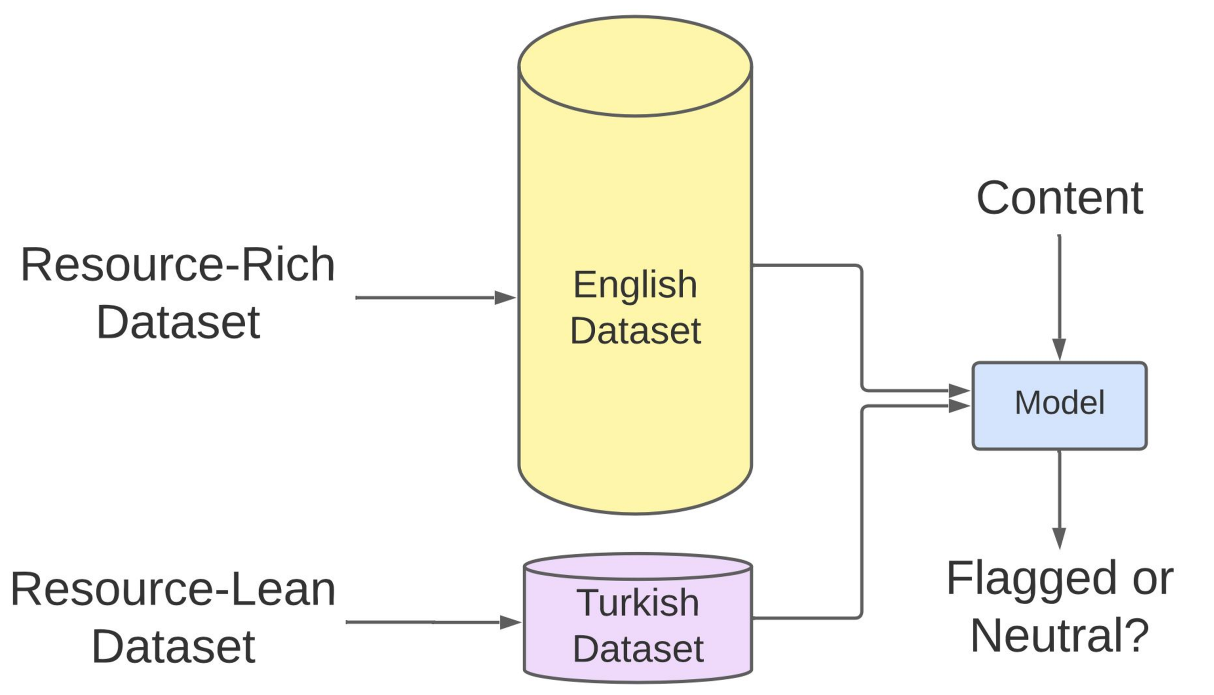
Sheikh Muhammad Sarwar[1,2], Dimitrina Zlatkova[3], Momchil Hardalov[3,4], Yoan Dinkov[3], Isabelle Augenstein[3,5], Preslav Nakov[3,6]

[1]Amazon.com, [2]University of Massachusetts, Amherst, [3]Checkstep Research
[4]Sofia University "St. Kliment Ohridski", [5]University of Copenhagen,
[6]Qatar Computing Research Institute, HBKU

Cross-Lingual Abusive Language Flagging

- Online abusive language harms users of online platforms and has the potential to incite violence [Muller and Schwarz, 2018].
- Types of abusive language that online platforms want to flag:
  - Hate speech
  - Offensive language
  - Cyberbullying
  - Hostile flames
  - Vulgar language
  - Insults
  - Profanity
  - …
- Inflammatory content on FB was up 300% before Delhi Riots
  - The New York Times had said that of India's 22 officially recognised languages, Facebook has trained its AI systems on five. But in Hindi and Bengali, it still did not have enough data to adequately police the content, and much of the content targeting Muslims "is never flagged or actioned."[1]

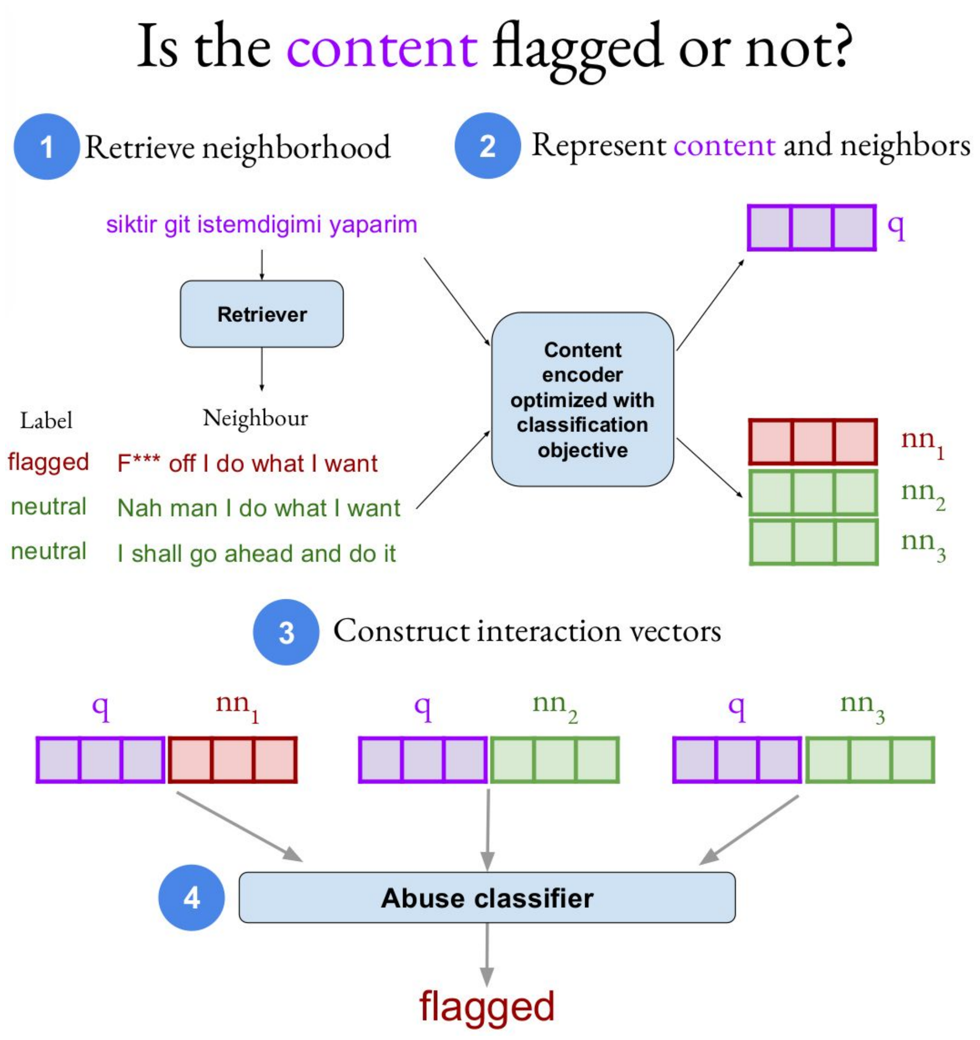
Problem Definition



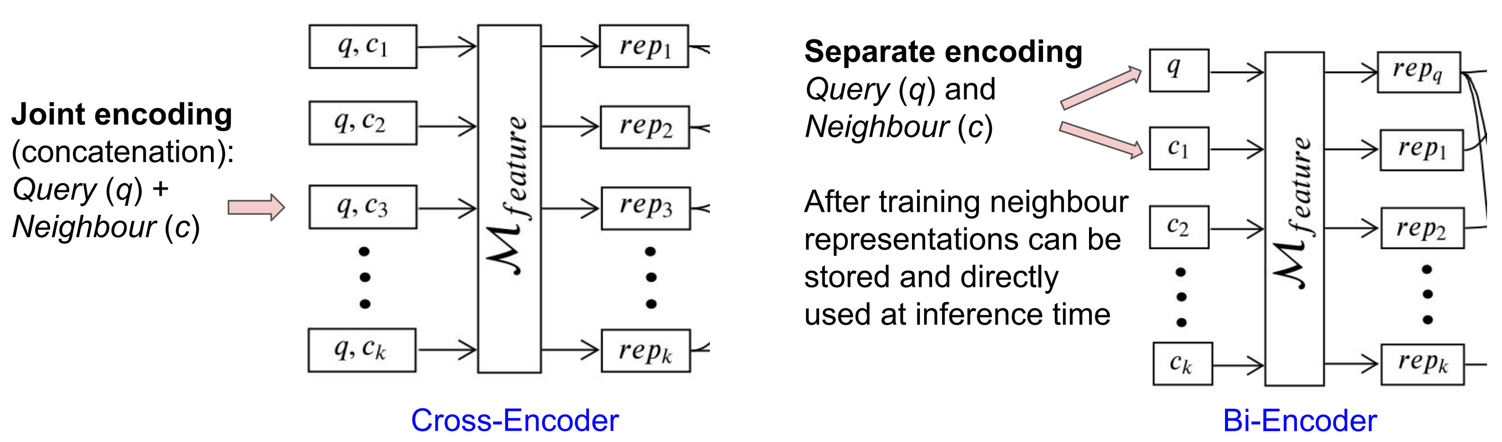
- Content flagging with two classes
- Cross-lingual transfer learning challenge

Our Proposed Framework (kNN+)

- Interaction vectors (core contribution)
- No explicit voting
- Understanding the neighbourhood at representation level

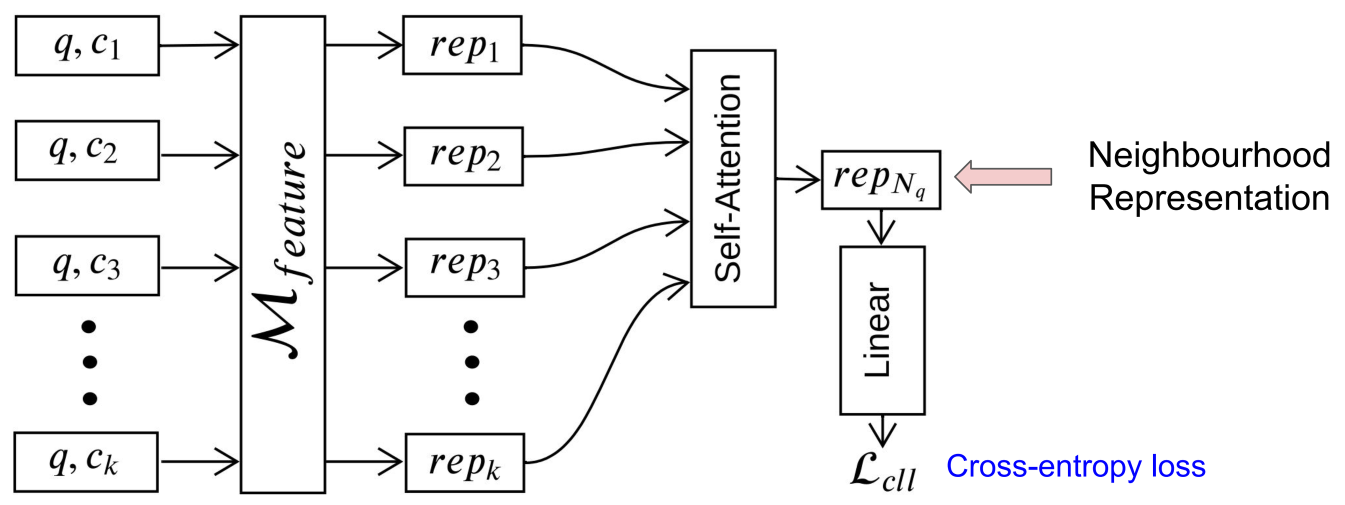


Query–Neighbor Interactions

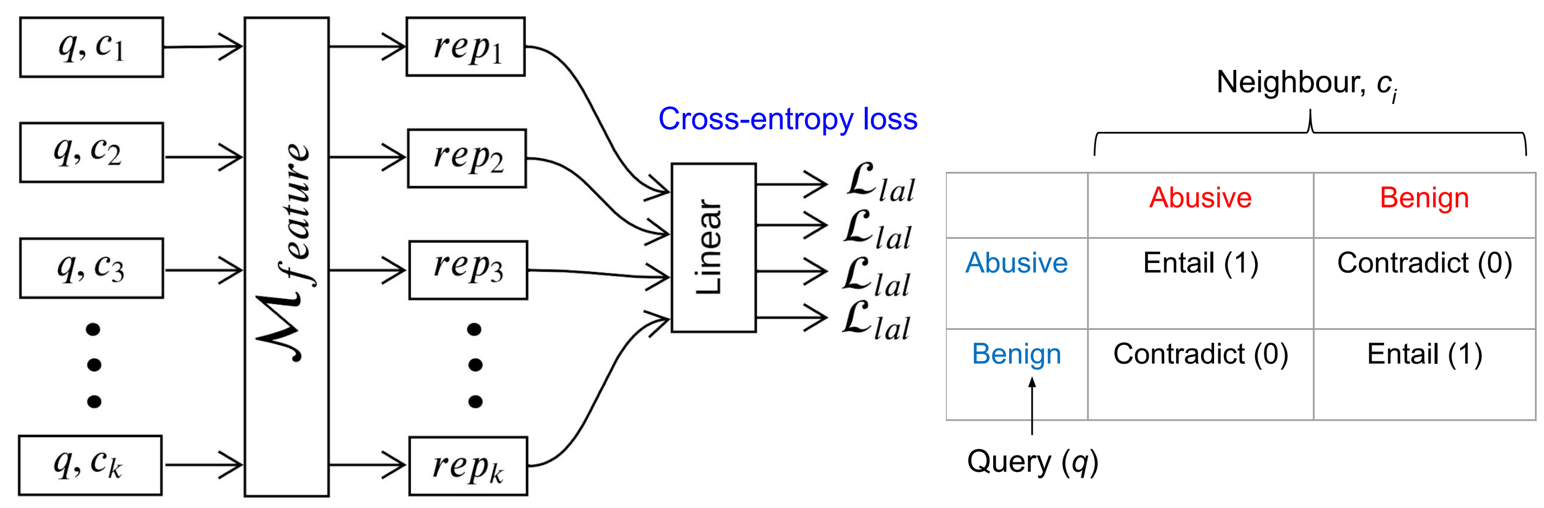


- *Two choices for $M_{feature}$*
  - **XLM-R** (base model)
  - **P-XLM-R** – XLM-R trained with paraphrastic knowledge based on a large number of paraphrases

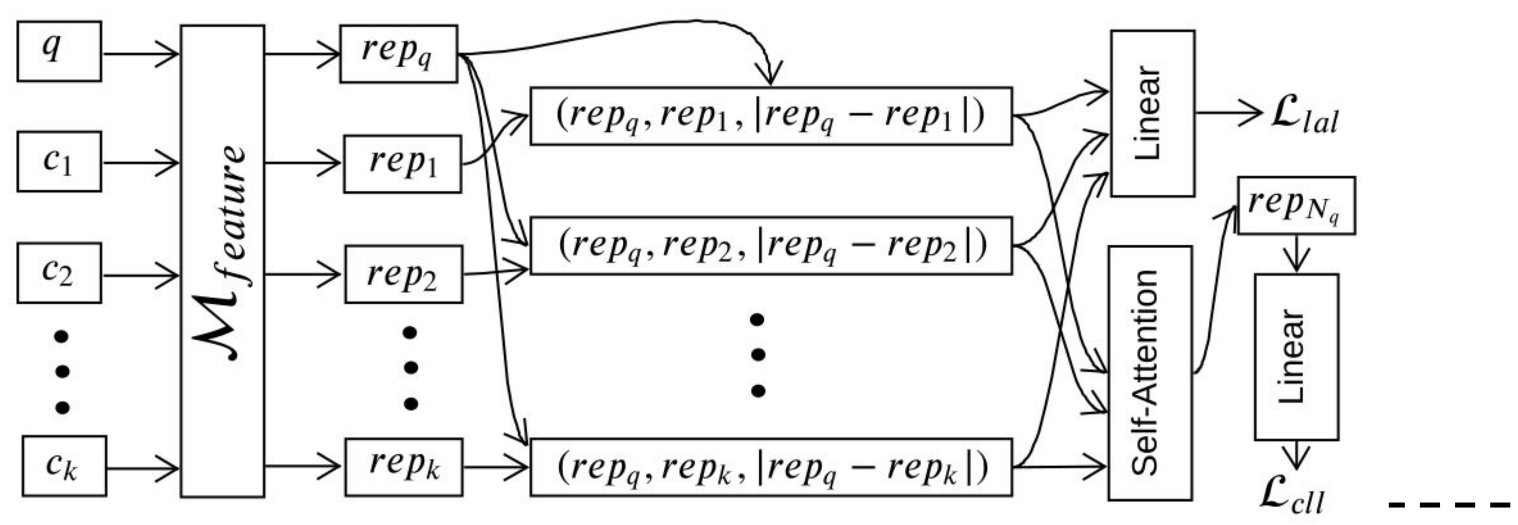
Cross-Encoder Architecture



Decision: *Is the query q Flagged or Neutral*
Classify *q* based on neighbourhood representation.
We know the label of *q* at training time.



Bi-Encoder Architecture



$$\mathcal{L} = (1-\lambda) \times \mathcal{L}_{lal} + \lambda \times \mathcal{L}_{cll}$$

Multi-task Learning Loss

Result: Cross-Lingual Transfer Learning

#	Method	Jigsaw Multilingual ES	IT	TR	WUL DE	EN	HR	RU	SQ	TR
1	Lexicon	35.8	40.5	34.0	70.9	70.6	63.9	63.6	58.2	71.8
2	FastText	55.3	47.2	64.2	74.2	72.7	58.9	74.2	65.9	72.5
3	XLM-R Target	<u>63.5</u>	56.4	80.6	82.1	75.7	73.2	76.7	77.3	78.8
4	XLM-R Mix-Adapt	**64.2**	58.5	76.1	83.2	**93.9**	87.3	82.1	86.2	86.0
5	XLM-R Seq-Adapt	60.5	58.3	81.2	83.9	88.0	80.0	80.0	86.3	83.5
6	LaBSE-kNN	44.7	48.5	66.0	70.8	77.1	84.1	79.1	83.1	75.6
7	Weighted LaBSE-kNN	44.8	38.3	52.1	71.7	85.4	82.4	79.5	83.7	81.0
8	CE kNN$^+$ + $\mathcal{M}^{\text{XLM-R}}_{feature}$	58.9	<u>63.8</u>	78.5	80.4	83.8	86.2	77.6	83.5	85.4
9	CE kNN$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature}$	59.4	**67.0**	<u>84.4</u>	84.8	88.0	86.3	83.8	83.0	86.5
10	CE kNN$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature}$ → SRC	61.2	61.1	**85.0**	**89.5**	<u>92.3</u>	**90.6**	84.9	<u>89.5</u>	<u>87.3</u>
11	BE kNN$^+$ + $\mathcal{M}^{\text{XLM-R}}_{feature}$	52.2	60.3	75.0	81.6	80.8	77.9	78.0	79.6	79.6
12	BE kNN$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature}$	58.8	56.6	80.6	83.8	86.9	82.2	**86.9**	84.9	83.7
13	BE kNN$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature}$ → SRC	59.1	59.5	81.6	<u>88.7</u>	90.7	<u>87.6</u>	<u>86.3</u>	**90.2**	**88.7**

* The feature extractor model could be ***XLM-R*** and ***P-XLM-R***[1].

** **SRC** indicates we pre-trained the neighbourhood model by using Jigsaw English as sources of query and neighbours.

Result: Examples of Nearest Neighbors

Turkish Query (flagged): siktir git istemdigimi yaparim

Text	BE kNN$^+$ Score	LaBSE Score	Label
[redacted] off i do what i want	0.99	0.88	flagged
[redacted] you i do what i want	1.0	0.84	flagged
i have going to do what ever i want	-0.19	0.83	neutral
[redacted] u i will do as i please	1.0	0.81	flagged
[redacted] off [redacted] off i do what i want [redacted]	1.0	0.77	flagged
nah man i do wat i want	-0.19	0.75	neutral
i shall go ahead and do it	-0.18	0.74	neutral
whaaat whateva i do what i want”	-0.2	0.72	neutral
ok i will do it	-0.17	0.69	neutral
great i will do what you are saying	-0.16	0.68	neutral

Re-rank items based on BE kNN+ scores and compute majority voting at rank 5

Conclusions

- Neighbourhood framework is effective for cross-lingual transfer learning
- Separate encoding of query and neighbours are effective
  - Possible to store dense vector representation of the neighbourhood database
  - Inference: retrieve neighbourhood and classify content
  - Neighbourhood database enrichment without re-training
- Explanation of classification decision based on influential neighbours