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

60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)

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
  - ...

Motivation: Lack of Multilingual Resources

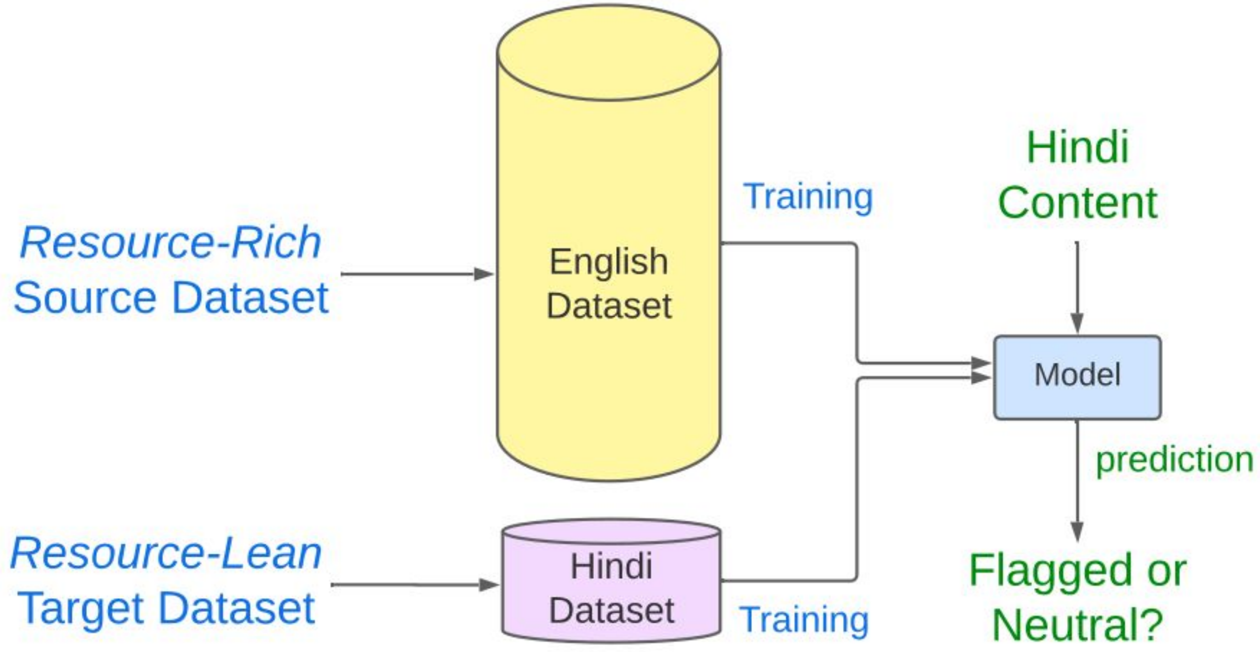
Inflammatory content on FB was up 300% before Delhi riots, says internal report

Internal Facebook documents accessed by US media showed that the company's researchers found Indian users are subject to "large amount of content that encourages conflict, hatred and violence".

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]

[1] https://www.thenewsminute.com/article/inflammatory-content-fb-was-300-delhi-riots-says-internal-report-156878

Problem Definition



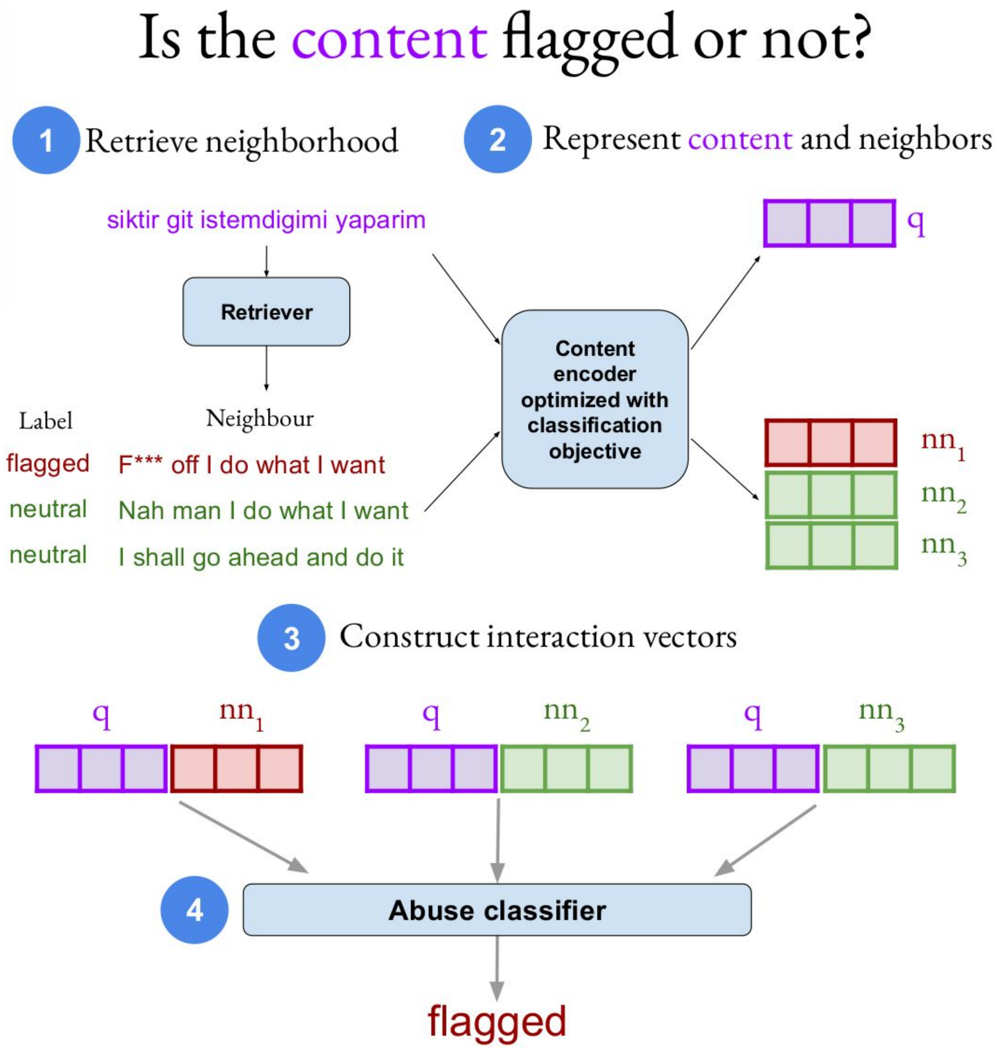
- Abusive language flagging with two classes
- Cross-lingual transfer learning challenge

Contributions

- Novel framework: *k*NN$^+$
  - cross-lingual transfer learning
    - *works with hundreds of labelled data* from the target language
  - *dense-vector representation* of the neighbourhood
  - voting strategy *learned* from the data
- Improvements of up to *9.5 F1 points* over strong baselines
  - with eight languages from two datasets

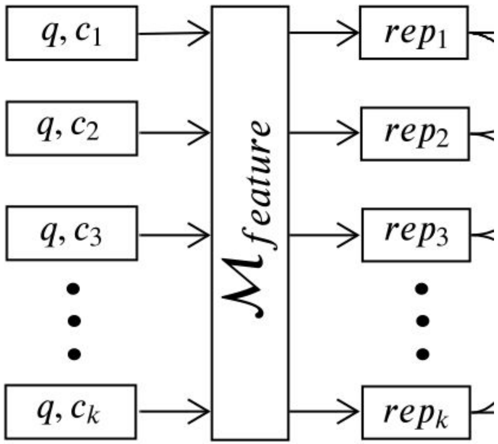
Our Proposed Framework (kNN+)

- Interaction vectors (core contribution)
- Voting is learned
- Neighbourhood representation using dense vectors



Query–Neighbor Interactions

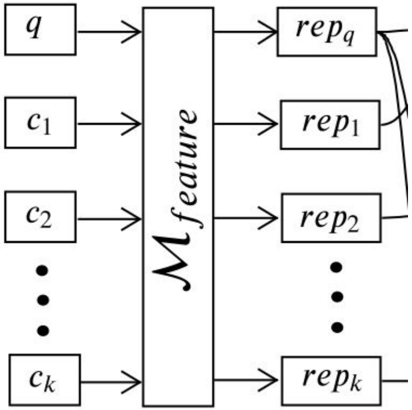
Joint encoding of *Query* (*q*) and *Neighbour* (*c*) by concatenation



Cross-Encoder

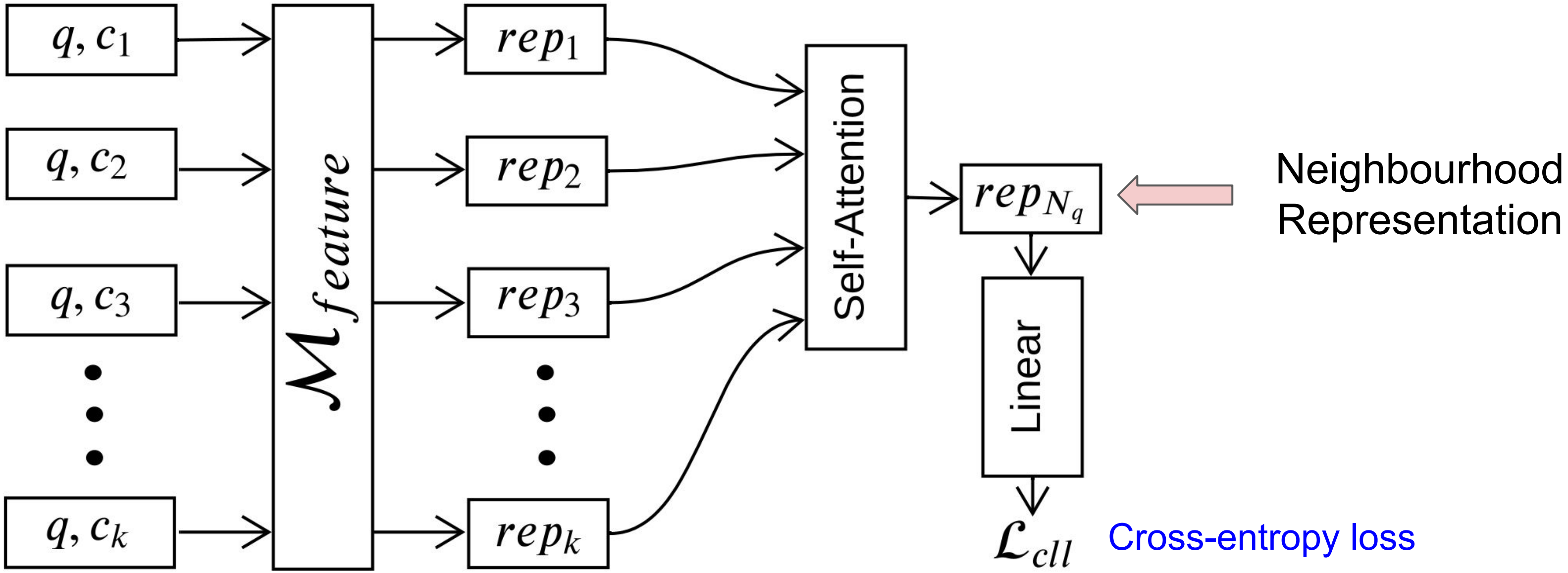
Separate encoding of *Query* (*q*) *and Neighbour* (c_i)

After training, neighbour representations can be stored and directly used at inference time



Bi-Encoder

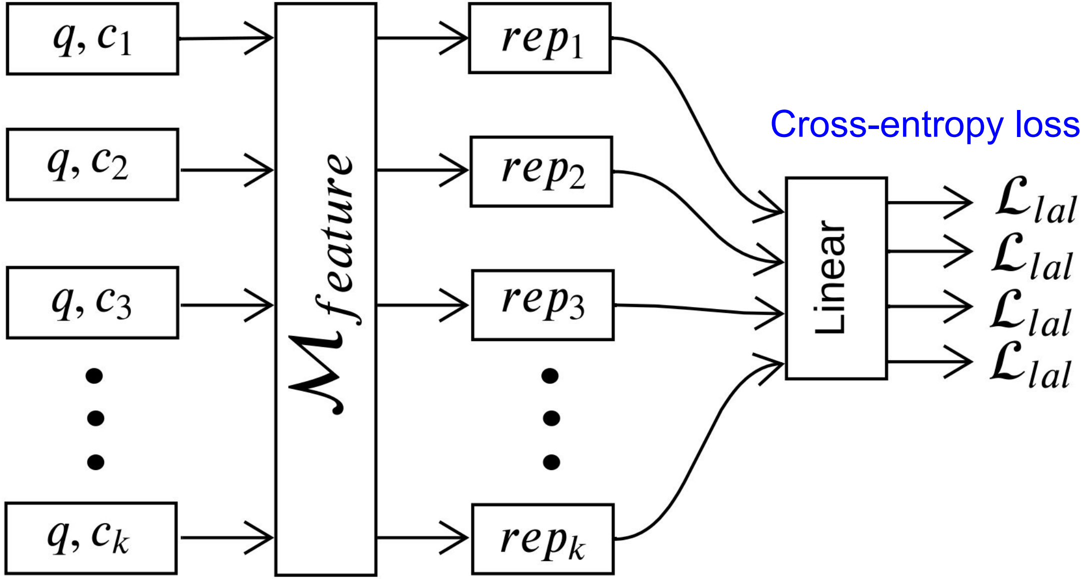
Cross-Encoder Architecture



Decision: *Is the query q Flagged or Neutral*

Classify *q* based on the neighbourhood representation.
We know the label of *q* at training time.

Cross-Encoder Architecture

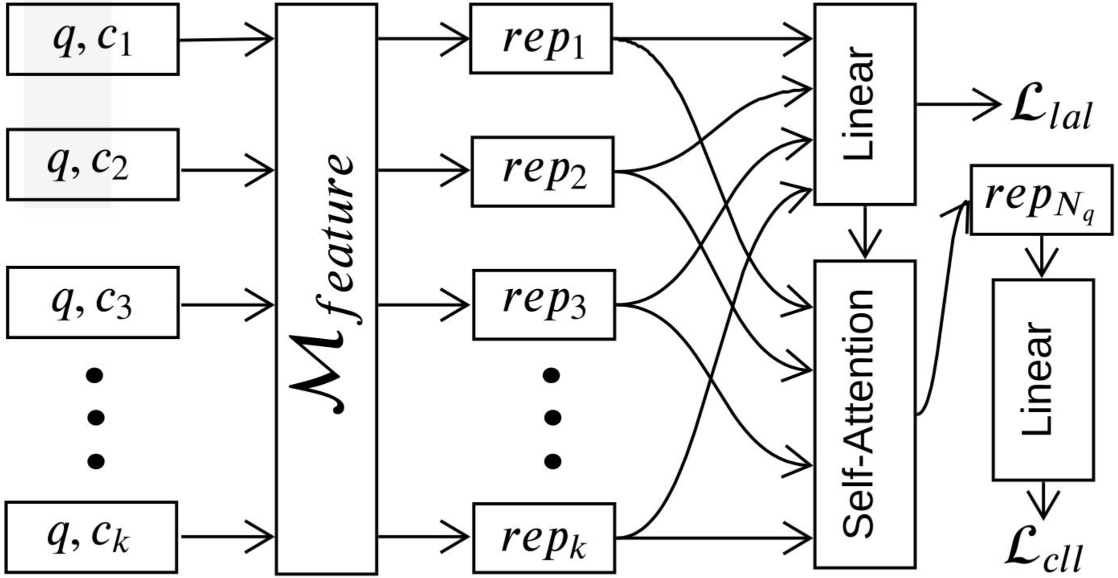


Neighbour (c_i)

	Flagged	Neutral
Flagged	Entail (1)	Contradict (0)
Neutral	Contradict (0)	Entail (1)

Query (q)

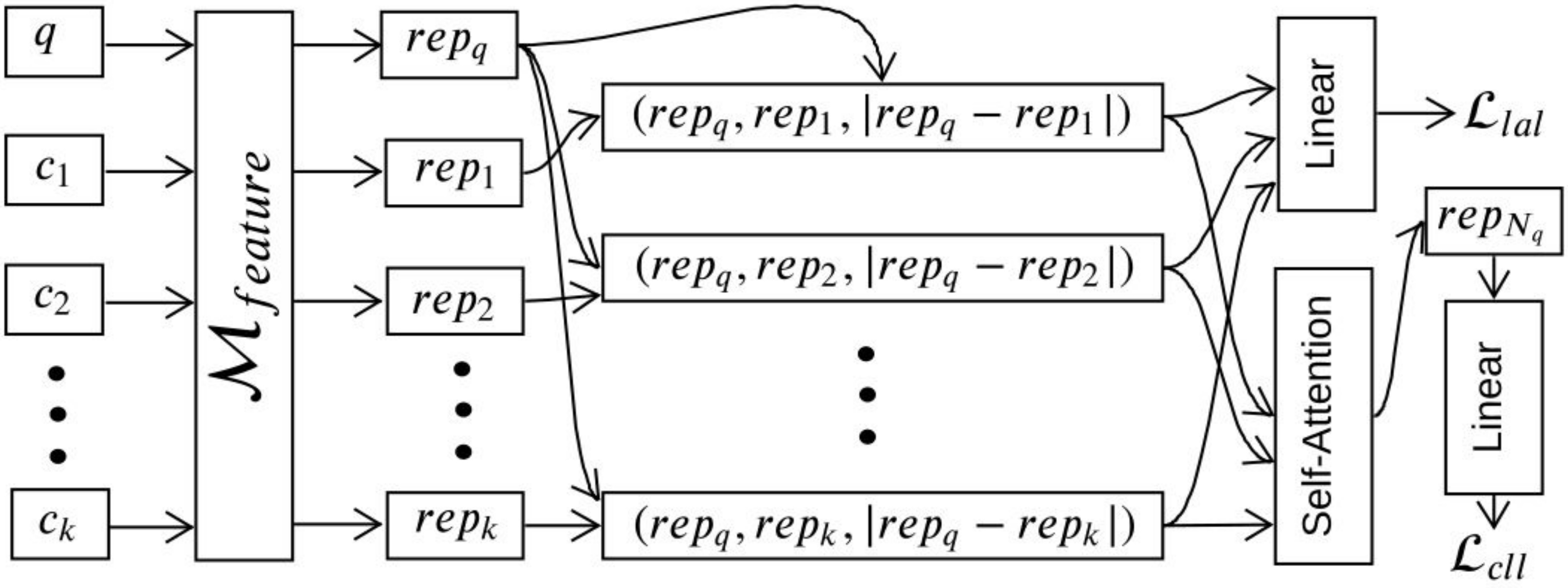
Cross-Encoder Architecture



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

Multi-task Learning Loss

Bi-Encoder Architecture

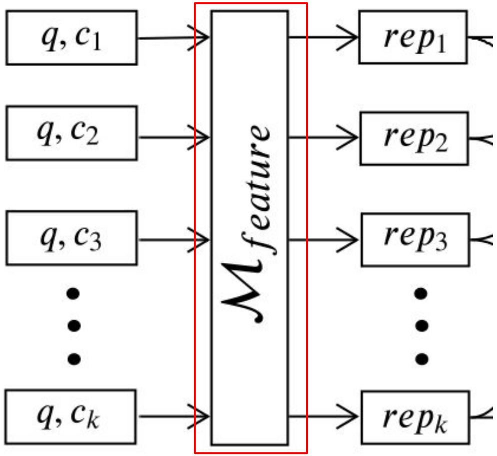


$$\mathcal{L} = (1 - \lambda) \times \mathcal{L}_{lal} + \lambda \times \mathcal{L}_{cll}$$

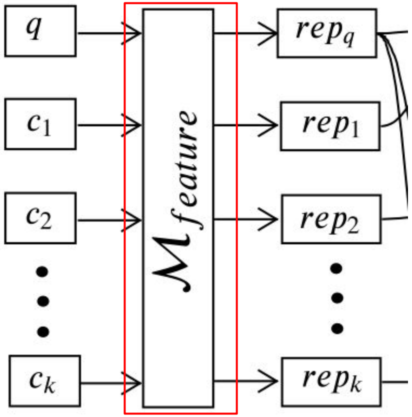
Multi-task Learning Loss

Choice of $M_{feature}$

- *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



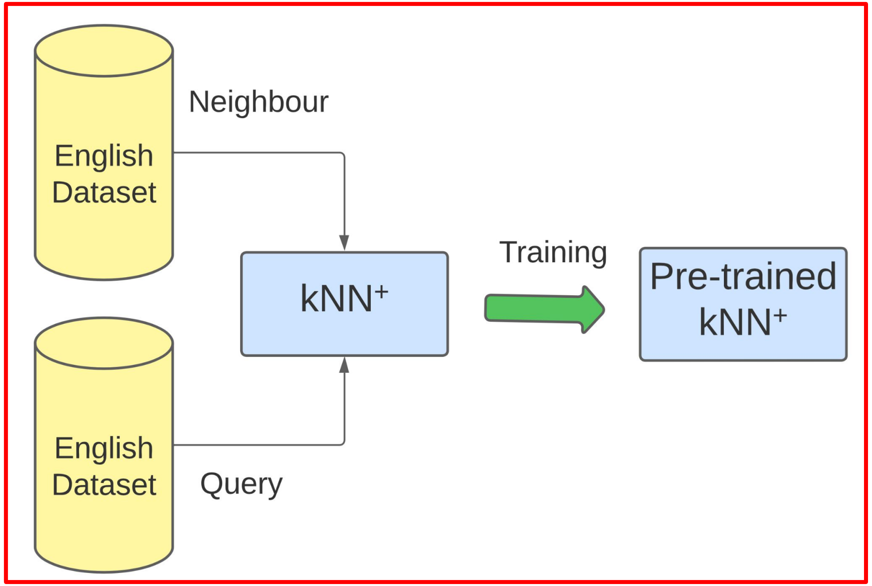
Cross-Encoder

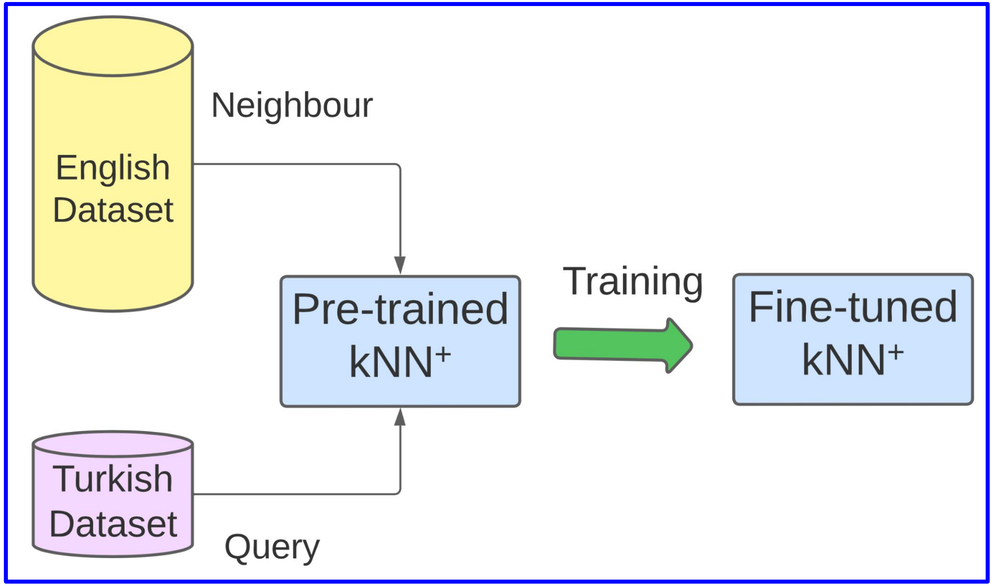


Bi-Encoder

Pre-training with Source Data (SRC)

We use the **resource-rich English** dataset for both *query* and *neighbours*.





Dataset Statistics

Jigsaw En(glish)

- **160K examples**
  (used for training only)
- Language: English (**EN**).

Jigsaw Multi(lingual)

- **8K examples**
  (500/600 for val/dev per language)
- Large class imbalance (**only 15% flagged**)
- Languages: Italian (**IT**), Turkish (**TR**), Spanish (**ES**).

WUL (translation based)

- **600 examples per language**
  (100 for val/dev)
- Languages: German (**DE**), Hungarian(**HR**), Albanian (**SQ**), Turkish (**TR**), Russian (**RU**).

Dataset	Examples	Flagged %	Neutral %
Jigsaw En	159,571	10.2	89.8
Jigsaw Multi	8,000	15.0	85.0
WUL	600	50.3	49.7

Result: Cross-Lingual Transfer Learning

#	Method	Jigsaw Multilingual			WUL					
		ES	IT	TR	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$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature} \rightarrow$ 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$^+$ + $\mathcal{M}^{\text{P-XLM-R}}_{feature} \rightarrow$ 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: ***XLM-R*** or ***P-XLM-R***[1]

** **SRC** indicates pre-training with a large English Jigsaw resource

[1] P-XLM-R comes with paraphrastic knowledge.

Performance in a Multilingual Scenario

Model	Representations	F1
Seq-Adapt	XLM-R	64.4
CE kNN^{+}	$\mathcal{M}^{\text{XLM-R}}_{feature}$	64.2
	$\mathcal{M}^{\text{P-XLM-R}}_{feature}$	62.8
	$\mathcal{M}^{\text{P-XLM-R}}_{feature} \rightarrow$ SRC	65.1
BE kNN^{+}	$\mathcal{M}^{\text{XLM-R}}_{feature}$	65.5
	$\mathcal{M}^{\text{P-XLM-R}}_{feature}$	63.7
	$\mathcal{M}^{\text{P-XLM-R}}_{feature} \rightarrow$ SRC	**67.6**

- Pre-training with a source English dataset is effective as bi-encoders are data-hungry
- BE kNN^{+} with paraphrastic representations is most effective: both in terms of efficiency and effectiveness

Conclusion and Future Work

- Our neighbourhood framework is effective for cross-lingual transfer learning
  - *3.6* and *2.14* absolute improvement in F1 over a strong baseline
  - *separate encoding* of query and neighbours is effective
    - *retrieve* neighbourhood and *classify* content
- Future work
  - add labeled English data *without re-training*
  - *explanation* of classification decisions based on neighbours

Thank You for Listening!

If you have more questions, please contact

smsarwar@cs.umass.edu