The University Of Sheffield.

NOTTINGHAM TRENT UNIVERSITY

ESWC2018 - 5 June 2018

Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network

Ziqi Zhang[1], David Robinson[2], Jonathan Tepper[2]

z.zhang@sheffield.ac.uk

@ziqizhang_zz

1. Information Retrieval Research Group, Information School, The University of Sheffield, UK

2. Nottingham Trent University, UK

The University Of Sheffield.

Outline

NOTTINGHAM TRENT UNIVERSITY

- The problem of hate speech
- Motivation and contributions
- Methodology
- Findings
- Conclusion

The Problem of Hate Speech

- **Significant increase of hate speech on the social media**
  - E.g., 80% young people in the EEA region have encountered hate speech online and 40% felt attacked or threatened [1]
  - Targeted individuals feel intimidated in their life
  - Leading to hate crime
- **Social media (e.g. Twitter) is increasingly criticised**
  - The anonymity and mobility
  - Insufficient content moderation
- **Legislation is taking shape (e.g., Germany), but it is very slow and challenging**

Motivation

- **To deal with it we must firstly know, how to identify hate speech?**
  - It is estimated that hundreds of millions are spent on manual content moderation annually [2]
  - But this simply does not scale up
- **Automated detection and understanding is crucial**
  - **Semantic content analysis** using NLP [3-14]
  - **Trends**: from feature engineering (as in 'classic' machine learning), to topology engineering (as in **D**eep **N**eural **N**etwork methods)
  - **Results are difficult to compare**: a mixture of public/non-public datasets, from different Web sources
- **The issues are**
  - We don't know which method is working better
  - Can we, and how can we do better

The University Of Sheffield.

Contribution

NOTTINGHAM TRENT UNIVERSITY

A new method based on Convolutional + Gated Recurrent Network (CNN + GRU, a type of DNN) that is shown to potentially **improve state of the art** on currently the **most complete collection of public Twitter datasets**

Method - Basic Principles

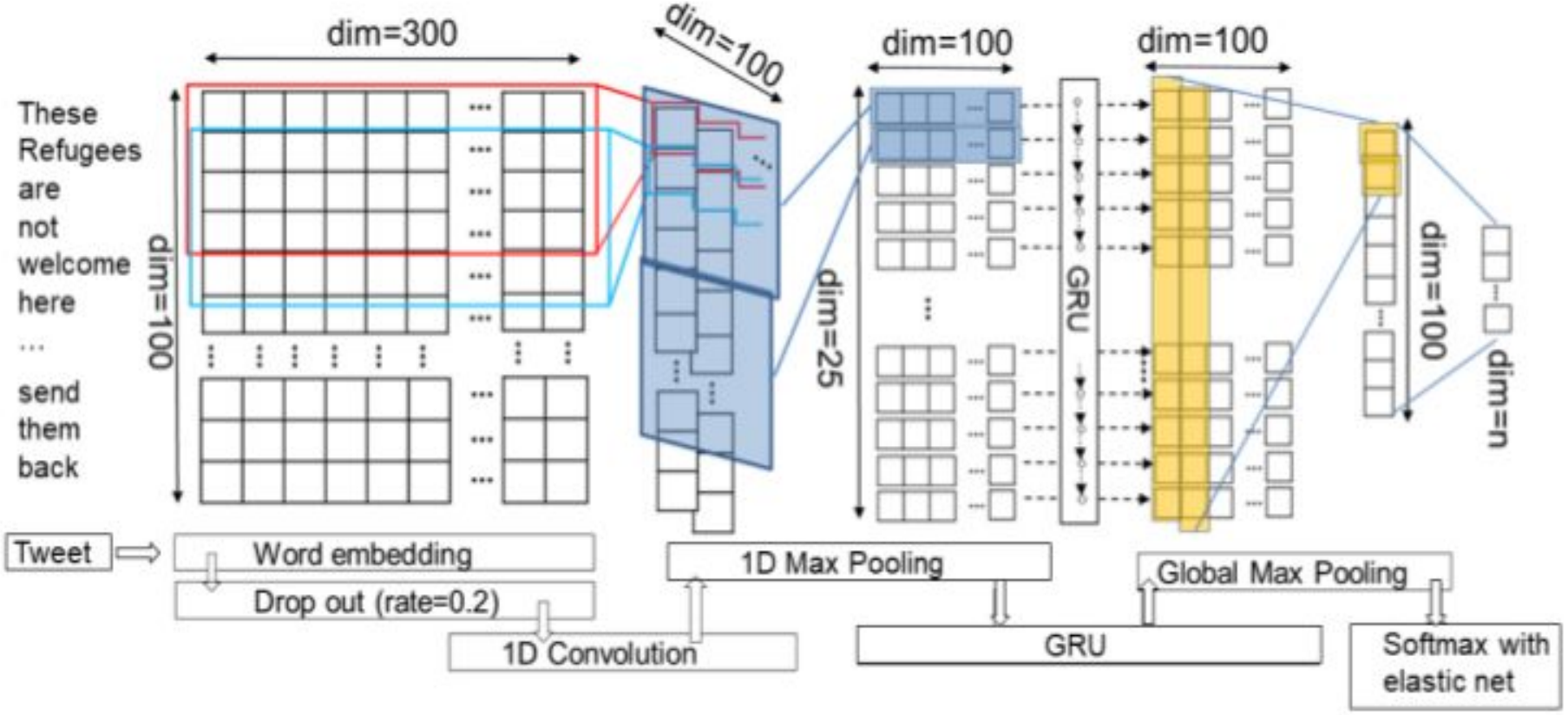
A typical NLP problem: short text classification using supervised machine learning

- **Classic methods focus on feature engineering**
  - SVM/Naive Bayes/Logistic regress etc.
  - Numerous feature types introduced
  - Poster (7 June): 'Hate Speech Detection on Twitter: Feature Engineering v.s. Feature Selection'
- **Deep learning based methods focus on topology engineering (**or, feature **learning)**
  - Word/character embedding + CNN (Convolutional Neural Network) or RNN (Recurrent Neural Network), typically LSTM[1]
  - Arguably works better than classic methods

1. LSTM: Long Short-Term Memory

Method - Our Take

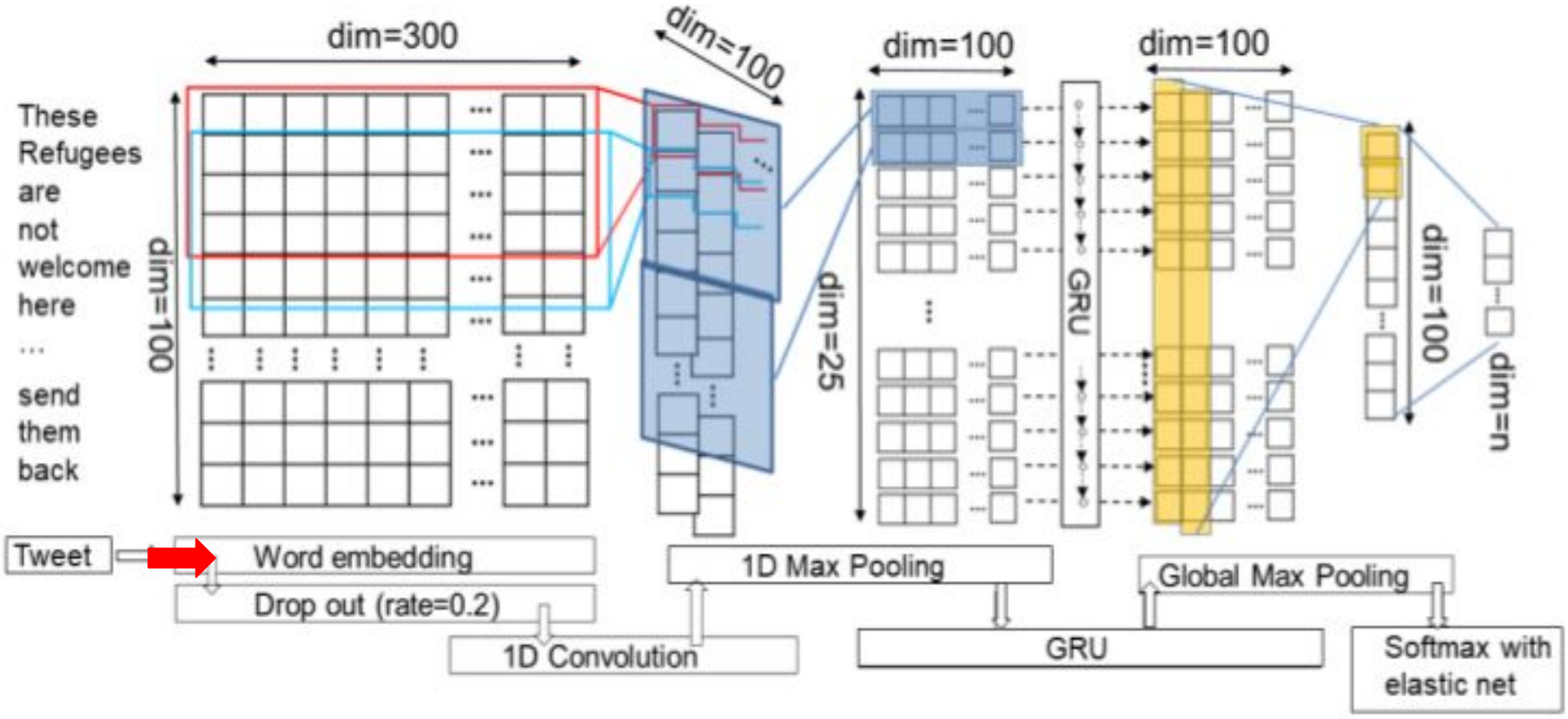
CNN + GRU (Gated Recurrent Unit): Overview



Method - Our Take

CNN + GRU: Word embedding

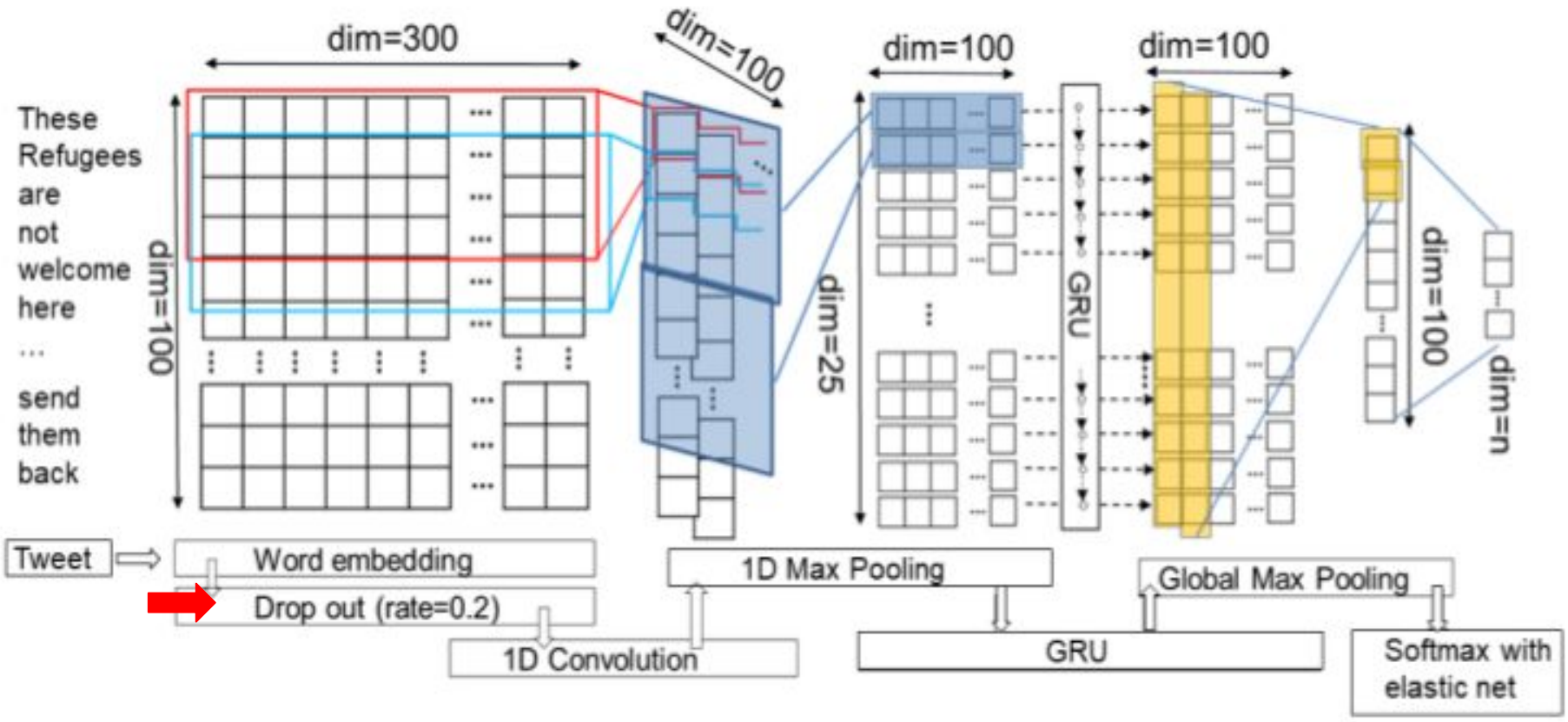
- Each word mapped to 300-dimensional real valued vector, using the pre-trained word2vec skip-gram
- Each Tweet as 100 words long truncating long messages and pad the shorter messages with zero values



Method - Our Take

CNN + GRU: Dropout

- Intuitively, this randomly removes a word in sentences and forcing the classification not to rely on any individual words, in order to address over-fitting

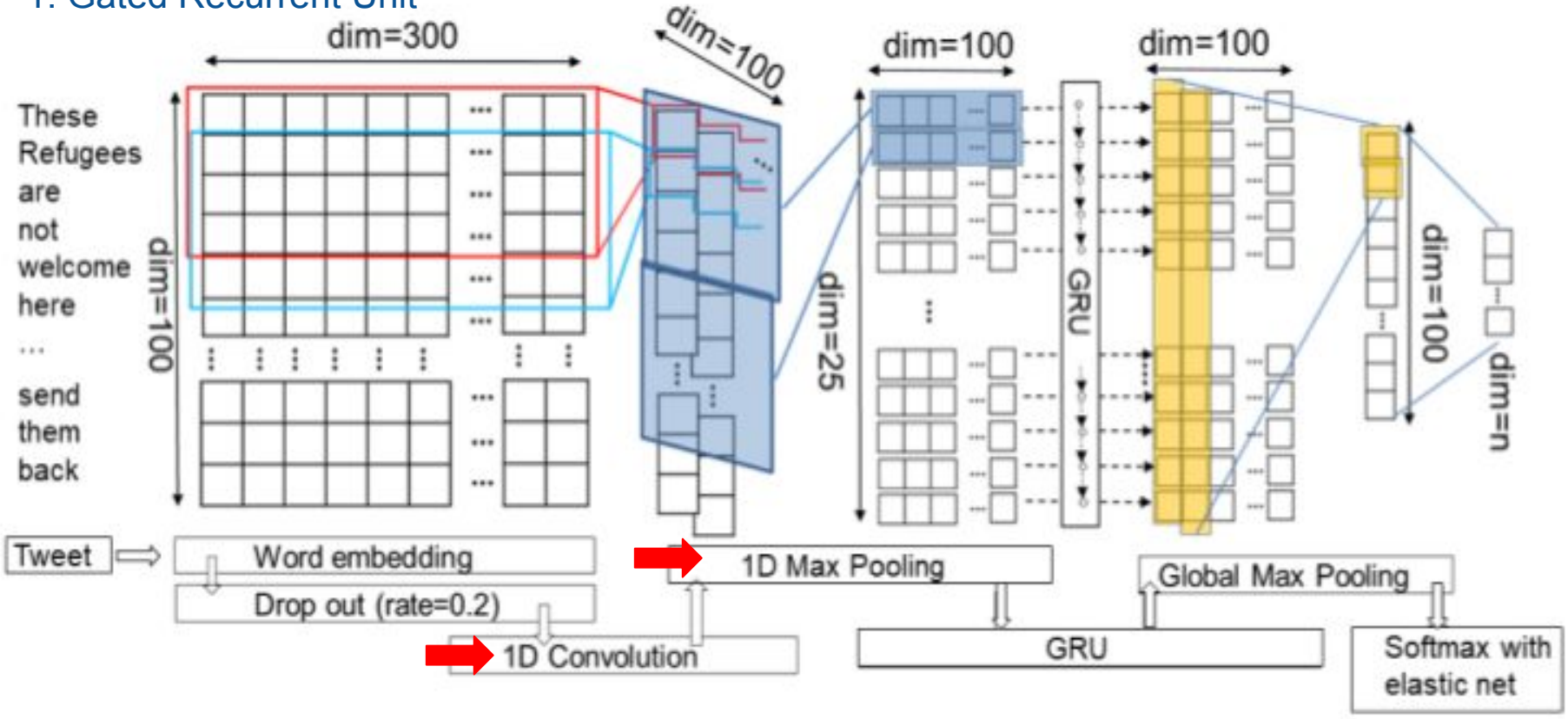


Method - Our Take

CNN + GRU[1]: CNN as feature extractors, intuitively:

- The convolutional layer extracts n-grams (weighted)
- The max-pooling layer selects most predictive n-grams
- See the paper for parameter details

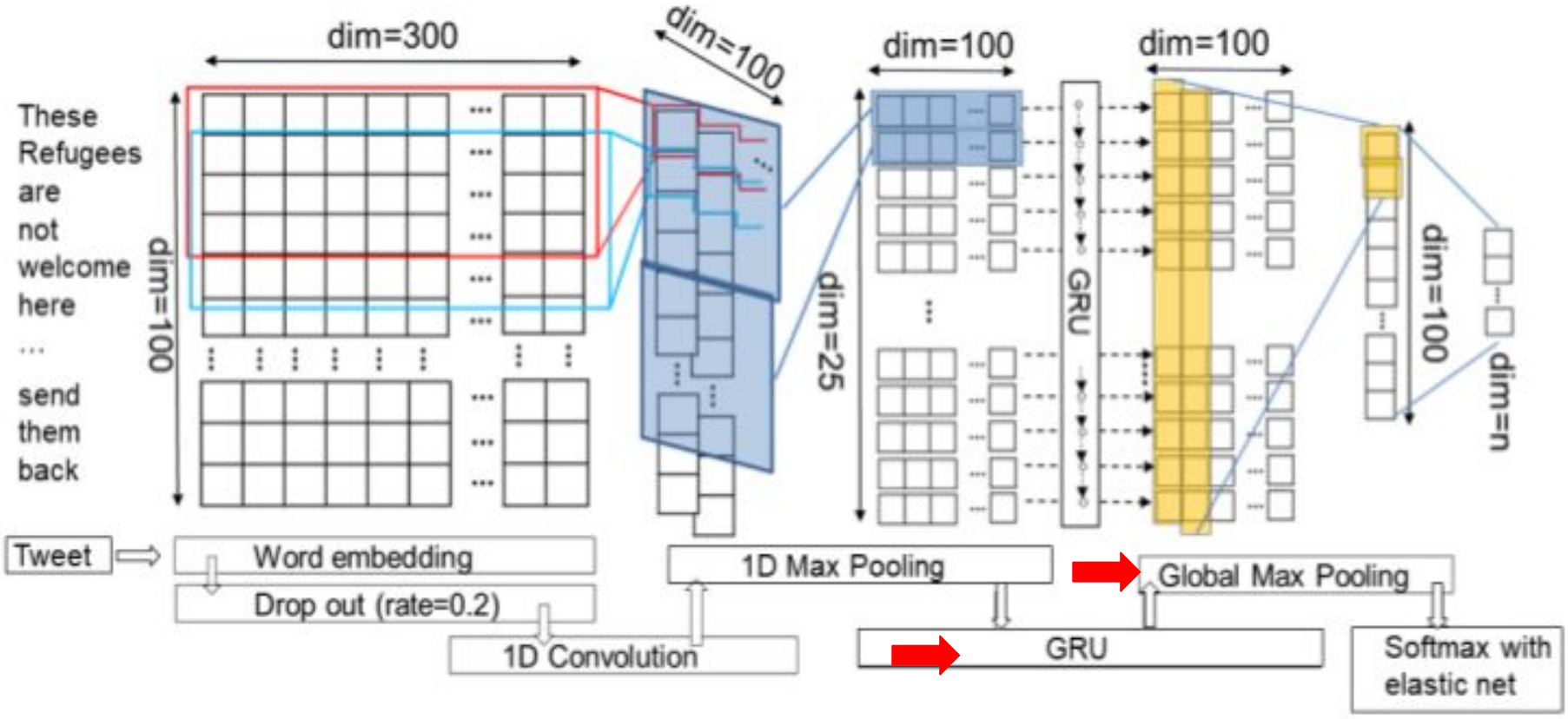
1. Gated Recurrent Unit



Method - Our Take

CNN + GRU: GRU

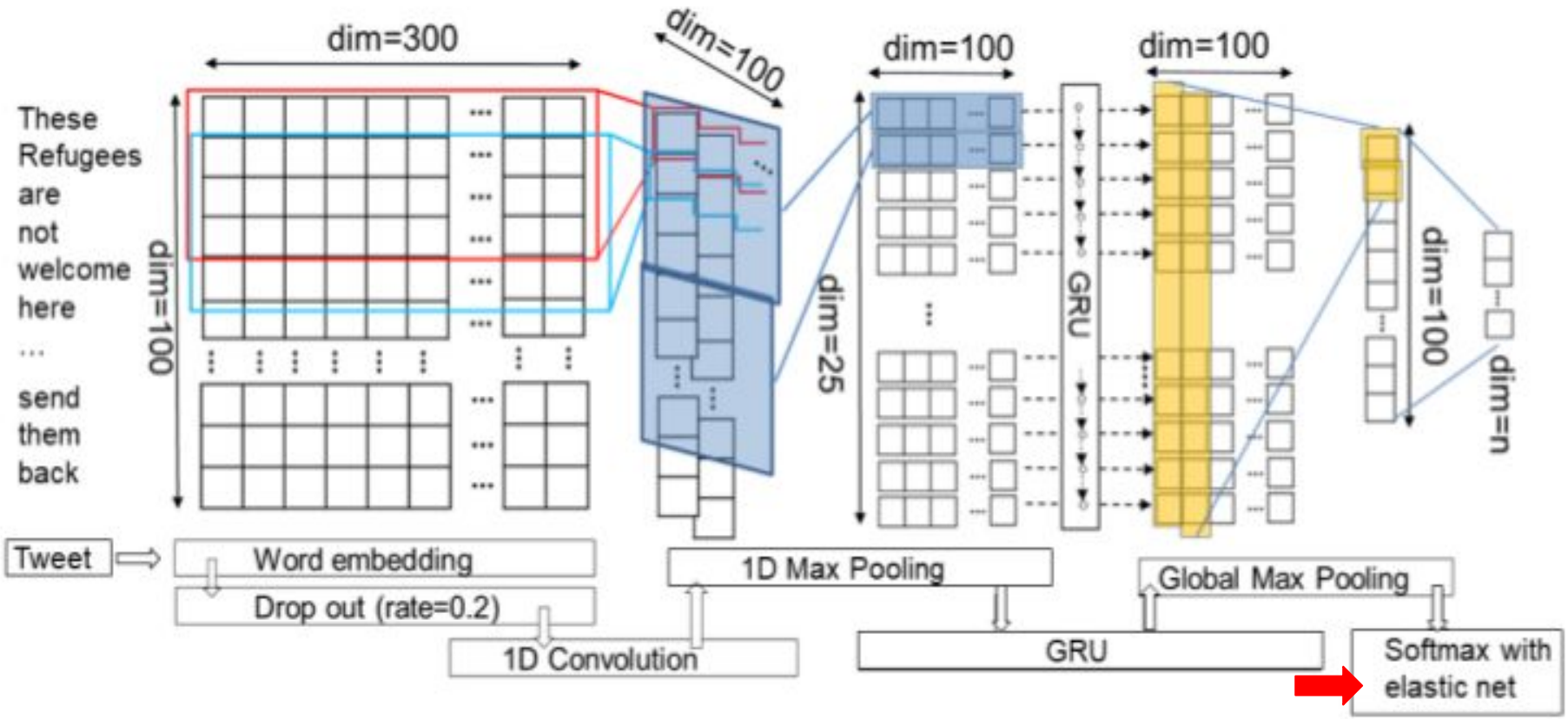
- Capture co-occurring n-grams and their order, e.g., as in pairs (refugees, be deported) and (refugees, not welcome) in the sentence ‘These refugees are not welcome in my Country they should all be deported ...’.



Method - Our Take

CNN + GRU: Softmax with elastic net regularisation

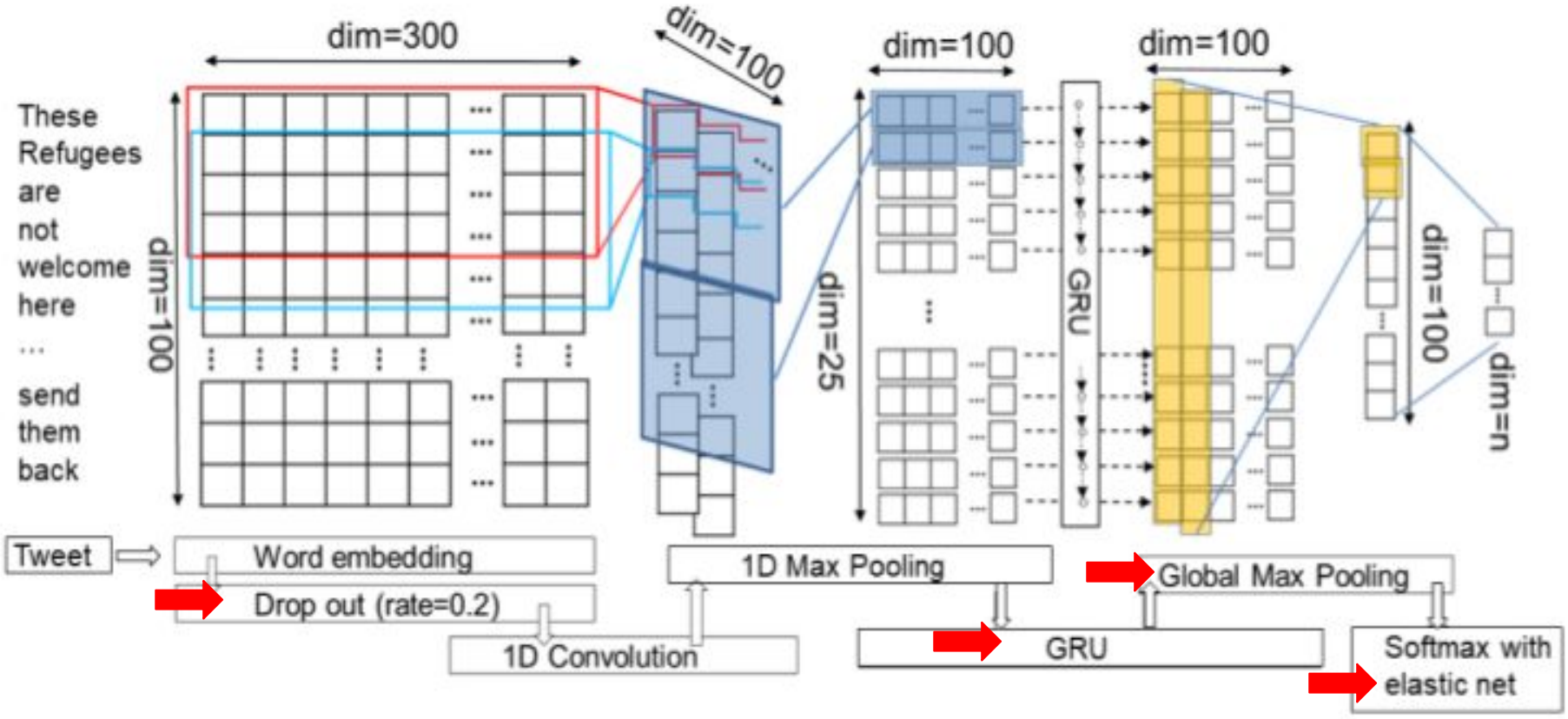
- Softmax predicts probability distribution over target labels
- Elastic net (combining L_1 and L_2 norms) regularisation to address overfitting



Method - Our Take

CNN + GRU: Summary of novelty

- v.s. DNN methods for hate speech detection: **combining CNN + GRU** to extract new features
- v.s. similar DNNs in non-language tasks: GRU with global max pooling for simpler structure, and regularisation tweaks



Findings

Dataset

- A total of 7 publicly available Twitter datasets
- So far the most comprehensive collection of hate speech datasets used in any studies

Dataset	#Tweets	Classes (#tweets)	Targeting characteristics
WZ-L	16,093	racism (1,934) sexism (3,149) neither (11,010)	racism, sexism
WZ-S.amt	6,594	racism (123) sexism (1,073) both (15) neither (5,383)	racism, sexism
WZ-S.exp	6,594	racism (85) sexism (777) both (35) neither (5,697)	racism, sexism
WZ-S.gb	6,594	racism (90) sexism (911) both (27) neither (5,564)	racism, sexism
WZ-LS	18,625	racism (2,012) sexism (3,769) both (30) neither (12,810)	racism, sexism
DT	24,783	hate (1,430) non-hate (23,353)	general
RM	2,435	hate (414) non-hate (2,021)	refugee, Muslim

Findings

Comparative models

Measure: micro-F1

- Baseline **SVM**: based on Davidson [4]
- Baseline **SVM+**: the above model adding 5 types of features
- Baseline **CNN**: CNN only, removing GRU.
- **CNN+GRU$_B$**: our model but removing regularisation tweaks (i.e., dropout, global max pooling and elastic net)
- Figures from previous studies on corresponding datasets

Dataset	SVM	SVM+	CNN	CNN+ GRU$_B$	CNN+ GRU	State of the art
WZ-L	0.74	0.74	0.80	0.81	**0.82**	0.74 Waseem [13] SVM
WZ-S.amt	0.86	0.87	0.91	**0.92**	**0.92**	0.84 Waseem [12] SVM
WZ-S.exp	0.89	0.90	0.90	0.91	**0.92**	0.91 Waseem [12] SVM
WZ-S.gb	0.86	0.87	0.91	0.92	**0.93**	0.90 Gamback [6] CNN
WZ-LS	0.72	0.73	0.81	0.81	0.82	0.82 Park [9], Word CNN 0.81 Park [9], Char CNN **0.83** Park [9], Hybrid
DT	0.87	0.89	**0.94**	**0.94**	**0.94**	0.87 SVM, Davidson [4]
RM	0.86	0.89	0.90	0.91	**0.92**	0.86 SVM, Davidson [4]

Findings

Error analysis (on 200 sample Tweets)

- Confusing features
  - ‘I’m a piece of white trash I say it proudly’
- Out of embedding vocabulary (OOV) words
  - ‘I’m just upset they got faggots on this show’
- Questioning or negation
  - ‘I honestly hate the term ‘feminazi’ so much. Stop it’
- Metaphors
  - ‘expecting gender equality is the same as genocide’
- Stereotypical views
  - ‘these same girls ... didn’t cook that well and aren’t very nice’

Findings

Language features may not be the ultimate answer

- I2U divides the number of Tweets of a class by the number of unique words found in that class - on average, how many Tweets share at least one word?
- U2C divides the number of words found only in a class by the number of words in that class (i.e., the words can be present in other classes at the same time).

	Racism		Sexism		Both		Non-hate		Hate	
	I2U	U2C	I2U	U2C	I2U	U2C	I2U	U2C	I2U	U2C
WZ-S.amt	.25	0	.84	0	.12	0	3.1	.22	-	-
WZ-S.exp	.26	.003	.70	.006	.19	0	3.3	.32	-	-
WZ-S.gb	.28	.003	.77	.008	.19	0	3.2	.29	-	-
WZ-LS	1.0	.004	1.4	.008	.16	0	3.8	.11	-	-
WZ-L	1.1	.004	1.3	.008	-	-	3.5	.11	-	-
DT	-	-	-	-	-	-	6.3	.45	.71	0
RM	-	-	-	-	-	-	2.2	.31	.65	.009

The University Of Sheffield.

Conclusion

NOTTINGHAM TRENT UNIVERSITY

- A new DNN based model for detecting hate speech on Twitter
- Outperforms state of the art in most cases on the largest collection of Twitter datasets
- The problem remains challenging due to the complexity in the language, where an ultimate answer may be difficult to find
- Future work
  - Other complex DNN structures for feature extraction
  - User-centric features, e.g., patterns in a user's language and their interactions

The University Of Sheffield.

References

NOTTINGHAM TRENT UNIVERSITY

1. EEA. Countering hate speech online, Last accessed: July 2017, http://eeagrants.org/News/2012/
2. B. Gambäck and U. K. Sikdar. Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online, pages 85–90. Association for Computational Linguistics, 2017.
3. P. Burnap and M. L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy and Internet, 7(2):223–242, 2015
4. T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detec- tion and the problem of offensive language. In Proceedings of the 11th Conference on Web and Social Media. AAAI, 2017.
5. N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidi- pati. Hate speech detection with comment embeddings. In Proceedings of the 24th International Conference on World Wide Web, pages 29–30. ACM, 2015.
6. B. Gambäck and U. K. Sikdar. Using convolutional neural networks to classify hate-speech. In Proceedings of the First Workshop on Abusive Language Online, pages 85–90. Association for Computational Linguistics, 2017
7. Y. Mehdad and J. Tetreault. Do characters abuse more than words? In Proceed- ings of the SIGDIAL 2016 Conference, pages 299–303, Los Angeles, USA, 2016. Association for Computational Linguistics.
8. C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive lan- guage detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, pages 145–153, 2016.
9. J. H. Park and P. Fung. One-step and two-step classification for abusive lan- guage detection on twitter. In ALW1: 1st Workshop on Abusive Language Online, Vancouver, Canada, 2017. Association for Computational Linguistics.
10. F. D. Vigna, A. Cimino, F. Dell'Orletta, M. Petrocchi, and M. Tesconi. Hate me, hate me not: Hate speech detection on facebook. In Proceedings of the First Italian Conference on Cybersecurity, pages 86–95, 2017.
11. W. Warner and J. Hirschberg. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, LSM '12, pages 19–26. Association for Computational Linguistics, 2012.
12. Z. Waseem. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In Proc. of the Workshop on NLP and Computational Social Science, pages 138–142. Association for Computational Linguistics, 2016.
13. Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL Student Research Workshop, pages 88–93. Association for Computational Linguistics, 2016.
14. G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In Conference on Information and Knowledge Management, pages 1980–1984. ACM, 2012.

The University Of Sheffield.

Thank you

NOTTINGHAM TRENT UNIVERSITY

ziqi.zhang@sheffield.ac.uk

@ziqizhang_zz