Kitami Institute of Technology

IJCAI-17

LaCATODA

LEARNING DEEP ON CYBERBULLYING IS ALWAYS BETTER THAN BRUTE FORCE

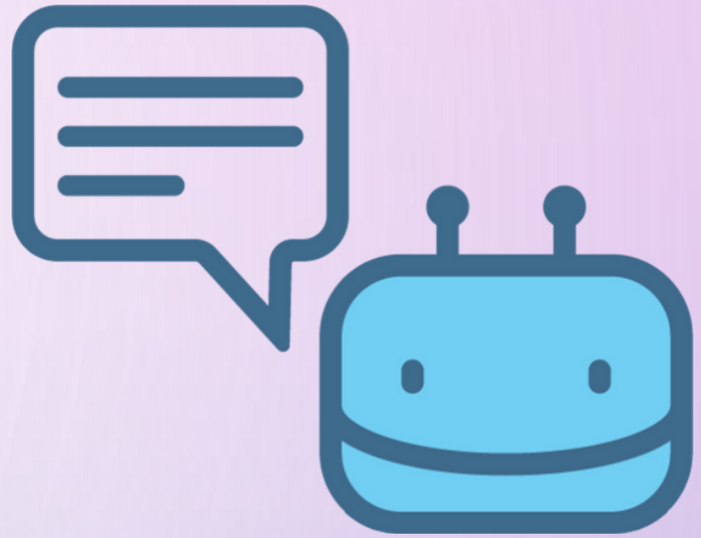
MICHAL PTASZYNSKI[1]

JUUSO KALEVI KRISTIAN ERONEN[2]

FUMITO MASUI[1]

1. KITAMI INSTITUTE OF TECHNOLOGY
2. TAMPERE UNIVERSITY OF TECHNOLOGY

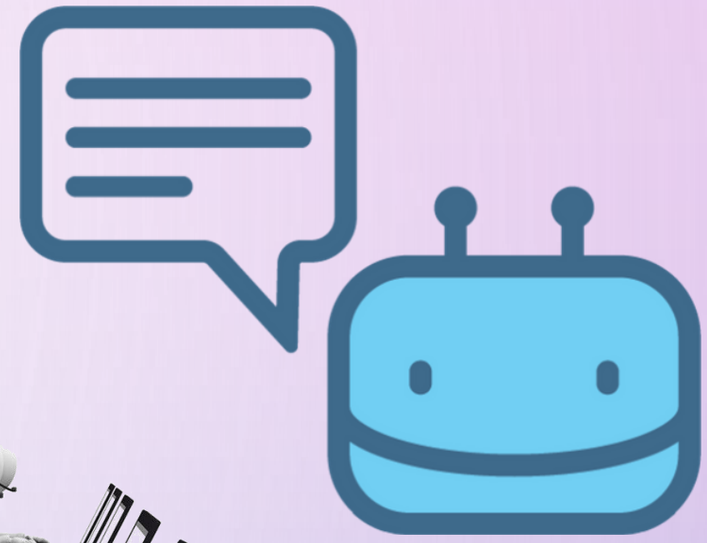
BACKGROUND



- DIALOG AGENTS APPLICATIONS

Q

BACKGROUND



- DIALOG AGENTS APPLICATIONS
  - CALL CENTERS
  - CUSTOMER SUPPORT
  - CASUAL CONVERSATIONS
  - LANGAUGE EDUCATION
  - ADVERTISEMENT / PROPAGANDA
  - PR / ANTI-PR
  - **FORUM MODERATION**

BACKGROUND

- FORUM MODERATION
  - ONE OF THE PROBLEMS ON FORA:

CYBERBULLYING

BACKGROUND

- CYBERBULLYING

" USING **TECHNOLOGY** TO RIDICULE OR **HUMILIATE** OTHERS "

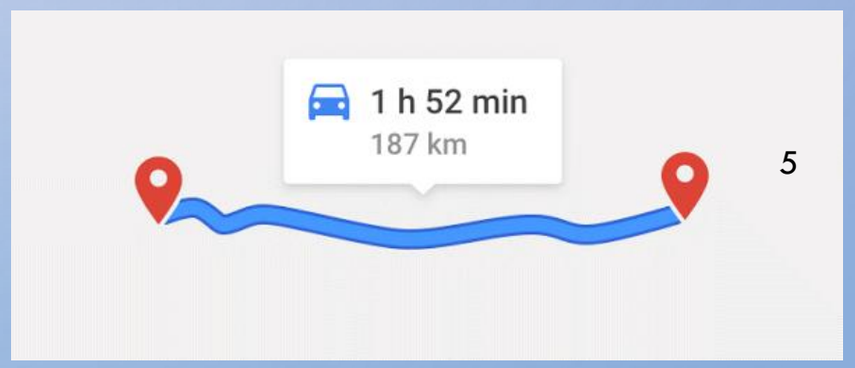
REPETITIVE

IMBALANCE OF POWER

(ONE VS MANY / WEAKER VS STRONGER)

EASY TO DO ON INETERNET

(LONGER PSYCHOLOGICAL DISTANCE)



BACKGROUND

- CYBERBULLYING

  CAUSES :

  AGGRESSION

  ALIENATION

  DEPRESSION

  SELF-MUTILATION

  SUICIDE

BACKGROUND

- CYBERBULLYING

CAUSES :

AGGRESSION

ALIENATION

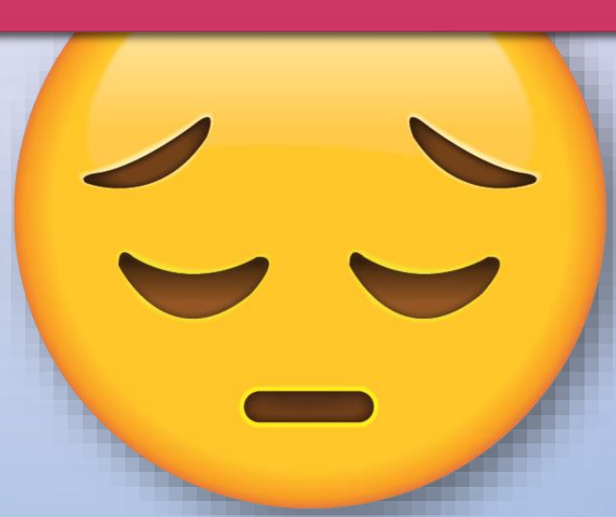
DEPRESSION

SELF-MUTILATION

SUICIDE

REAL WORLD PROBLEM

MORE LIFE ON INTERNET =
= MORE CYBERBULLYING



8% TO EVEN 20% OF USERS (OFTEN KIDS)

Q

BACKGROUND

- MANY INTERNET FORA (OVER 1 MIL.[1])

  IMPOSSIBLE TO MODERATE EVERYTHING MANUALLY

- ONLINE PATROL (TEACHERS, PARENTS VOLUNTEERS)
  - READ EVERYTHING TO FIND CYBERBULLYING
  - NOT ENOUGH TIME
  - PSYCHOLOGICAL BURDEN

NEED TECHNOLOGY SUPPORT

1) https://www.quora.com/How-many-online-forums-are-in-existence

PREVIOUS RESEARCH

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics = Brute Force Search

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Automatic acquisition of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

Affect analysis of cyberbullying data

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

SVM / optimization

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

SO-PMI-IR / phrases

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

2013 PATENT

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

Category Relevance Optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

Language Combinatorics = Brute Force Search

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Language Combinatorics / Preprocessing

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

Automatic acquisition of harmful words

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Language Combinatorics **= Brute Force Search**

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F.
R. Rzepka and K. Araki. 2010. **Machine Learn**
nd Affect Analysis Against Cyber-Bullying. I
oceedings of AISB'10, 29th March – 1st April 201

Category Relevance Optimization

Automatic acquisition of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

Language Combinatorics / Preprocessing

SO-PMI-IR / phrases

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. [illegible] K. Araki, and Y. Momouchi. 2010. **In the Service of Online: Tackling Cyber-Bullying with Machine Learning and [illegible] Analysis**. *International Journal of Computational Linguistics [illegible] earch*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics = Brute Force Search

SVM / optimization

[illegible]ta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. [illegible] 2013. **Detecting Cyberbullying Entries on Informal [illegible]l Websites Based on Category Relevance Maximi** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

[illegible]tegory Relevance [illegible]timization

Automatic acquisition of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics = Brute Force Search

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Automatic acquisition of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics
= Brute Force Search

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. **mproving Performa with Double Filtered** *mposium on Cloud Co*

nski, K. Yamamoto. 2015. **I llying Detection Method utual Information**. *ACM Sy (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Automatic **uisition** of harmfu **ords**

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics = Brute Force Search

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Affect analysis of cyberbullying data

Automatic acquisition of harmful words

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. A... **rute Force Works Best Against Bullying**, *IJCAI 2015 ... Intelligent Personalization (IP 2015)*, Buenos Aires, 201...

Language Combinator... **= Brute Force Searc...**

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski... ...oto. 2015. **Improving Performance of Cyberbullyi... ...n Method with Double Filtered Point-wise Mutual... ...on**. *ACM Symposium on Cloud Computing 2015 (SoC... ...ust 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Automatic acquisi... of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. **Study on the polarity classification model for the purpose of detecting harmful information on informal school sites** (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

Patent No. 2013-245813. Inventors: Fumito Masui, Michal Ptaszynski, Nitta Taisei. Patent name: *An Apparatus and Method for Detection of Harmful Entries on Internet*

M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Extracting Patterns of Harmful Expressions for Cyberbullying Detection**, *7th Language & Technology Conference (LTC'15)*, 2015.11.27-29.

SO-PMI-IR / phrases

Language Combinatorics / Preprocessing

2013 PATENT

Michal Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. 2015. **Brute Force Works Best Against Bullying**, *IJCAI 2015 Workshop on Intelligent Personalization (IP 2015)*, Buenos Aires, 2015.07.25-31

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, and Y. Momouchi. 2010. **In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis**. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154, 2010.

Language Combinatorics **= Brute Force Search**

SVM / optimization

T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. 2013. **Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization.** In *Proceedings of IJCNLP 2013*, pp. 579-586.

S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. 2015. **Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information**. *ACM Symposium on Cloud Computing 2015 (SoCC'15)*, August 2015.

Michal Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka and K. Araki. 2010. **Machine Learning and Affect Analysis Against Cyber-Bullying**. In *Proceedings of AISB'10*, 29th March – 1st April 2010.

Category Relevance Optimization

Automatic acquisition of harmful words

Affect analysis of cyberbullying data

2009 2010 2011 2012 2013 2014 2015

PREVIOUS RESEARCH

Feature sophistication

simple→ **→sophisticated**

bag-of-words

words

phrases

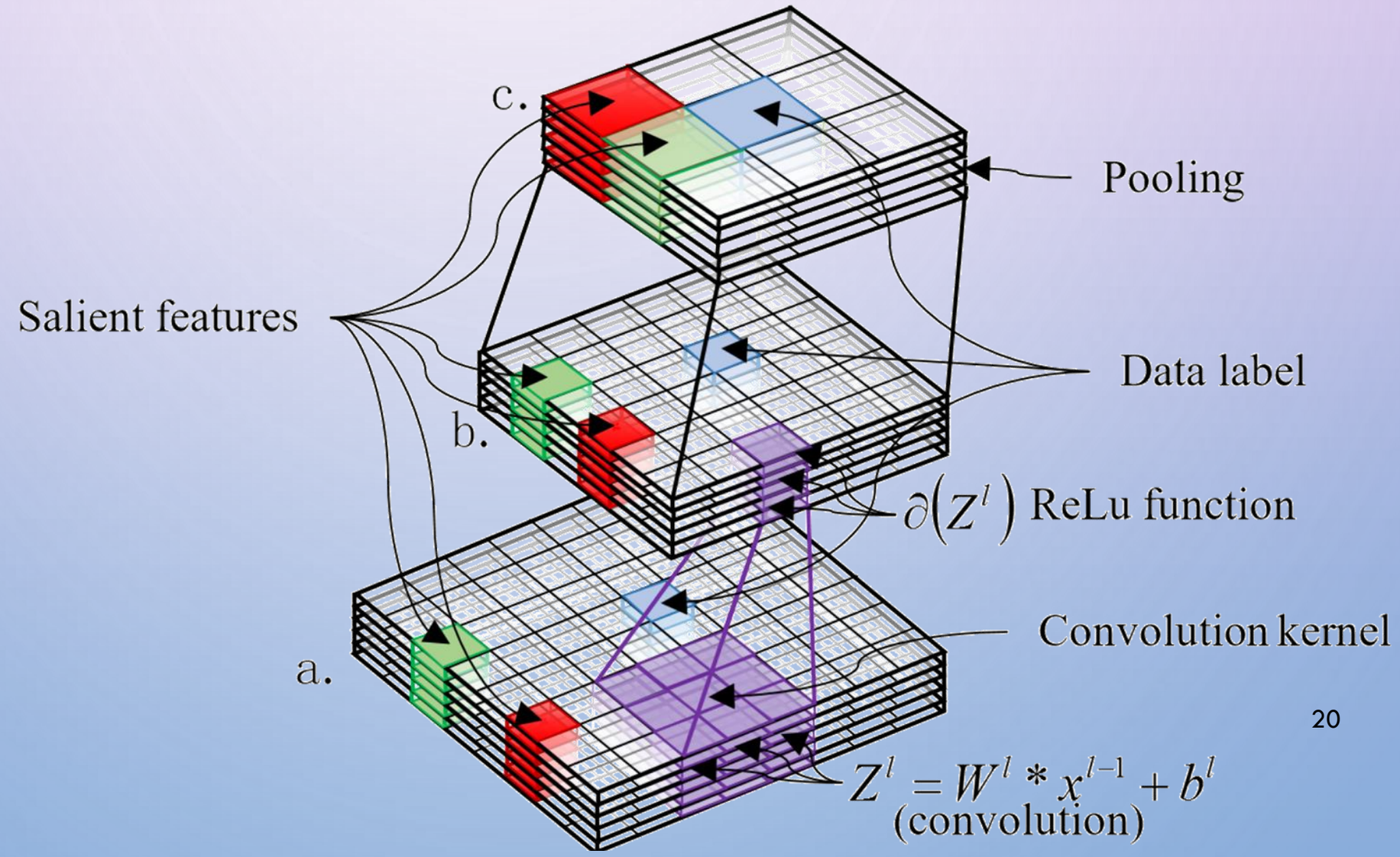
word patterns

syntactic pat.

DEEP LEARNING

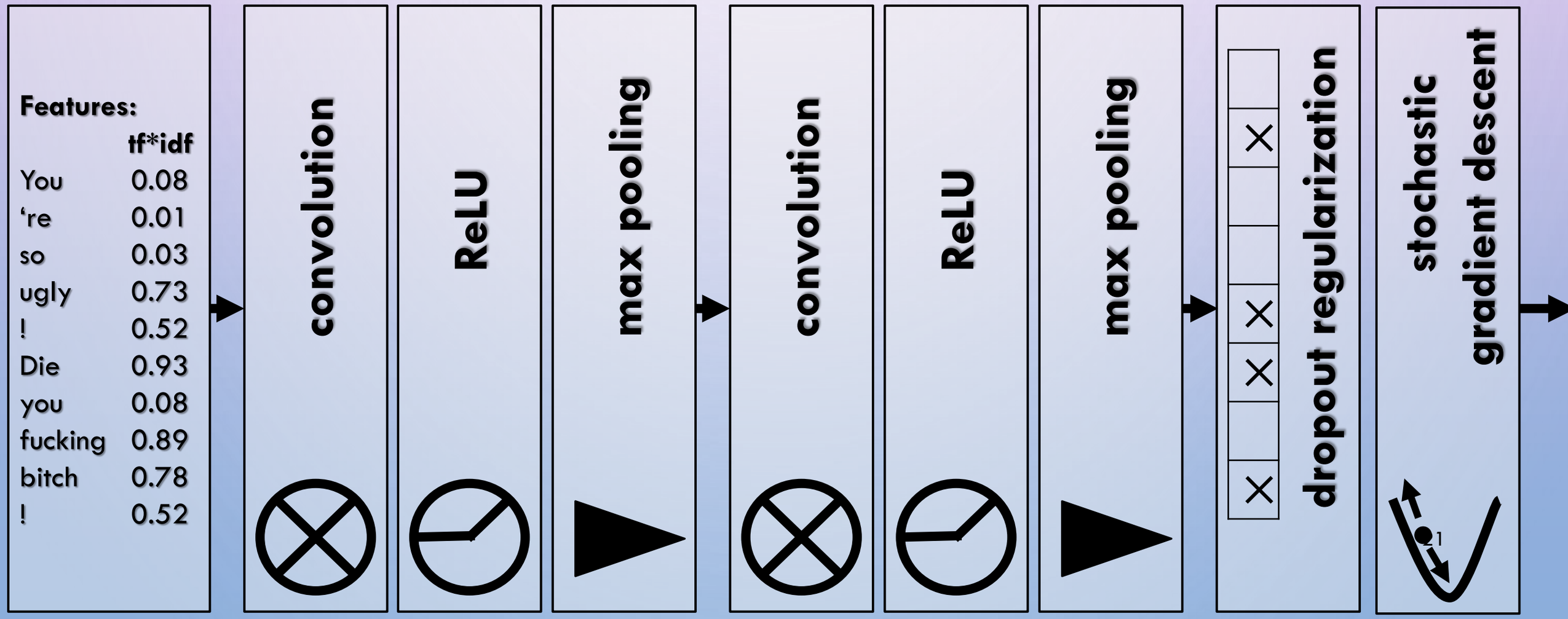
PROPOSED METHOD

- DEEP CONVOLUTIONAL NEURAL NETWORK



PROPOSED METHOD

- DEEP CONVOLUTIONAL NEURAL NETWORK



*) dummy weights only for explanation.

PROPOSED METHOD

- **CONVOLUTION**

Features:	**tf*idf**
You	0.08
‘re	0.01
so	0.03
ugly	0.73
!	0.52
Die	0.93
you	0.08
fucking	0.89
bitch	0.78
!	0.52

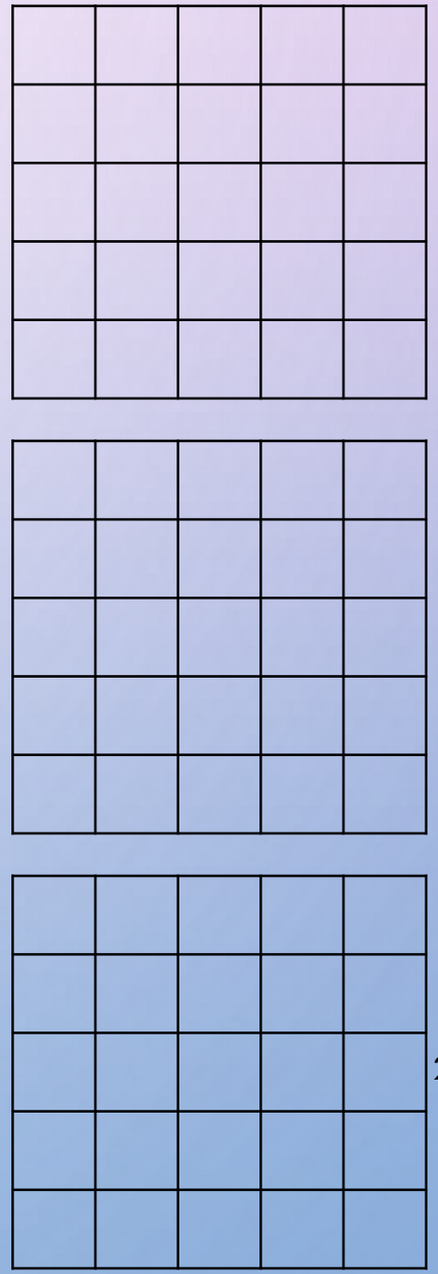
Batch = 5x5

You	‘re	so	ugly	!
Die	you	f*ng	b*ch	!
…				

Slide batch window

.08	.01	.03	.73	.52
.93	.08	.89	.78	.52
…				

Average of weights in batch for each feature



PROPOSED METHOD

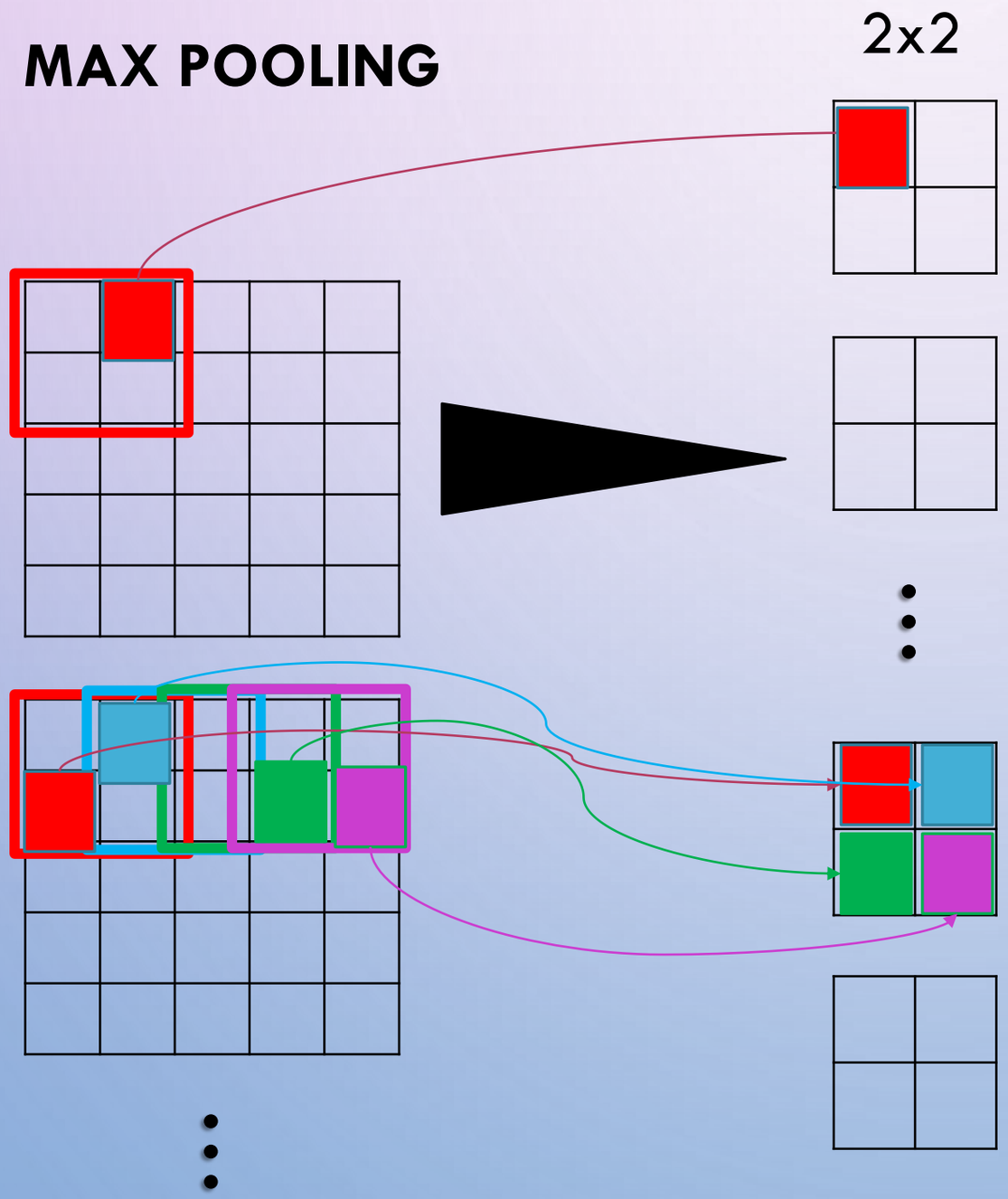
- **RECTIFIED LINEAR UNITS (RELU)**

.83	-.2	.04	.02	.1
.12	.76	-.32	.21	-.12
…				

.83	**0**	.04	.02	.1
.12	.76	**0**	.21	**0**
…				

PROPOSED METHOD

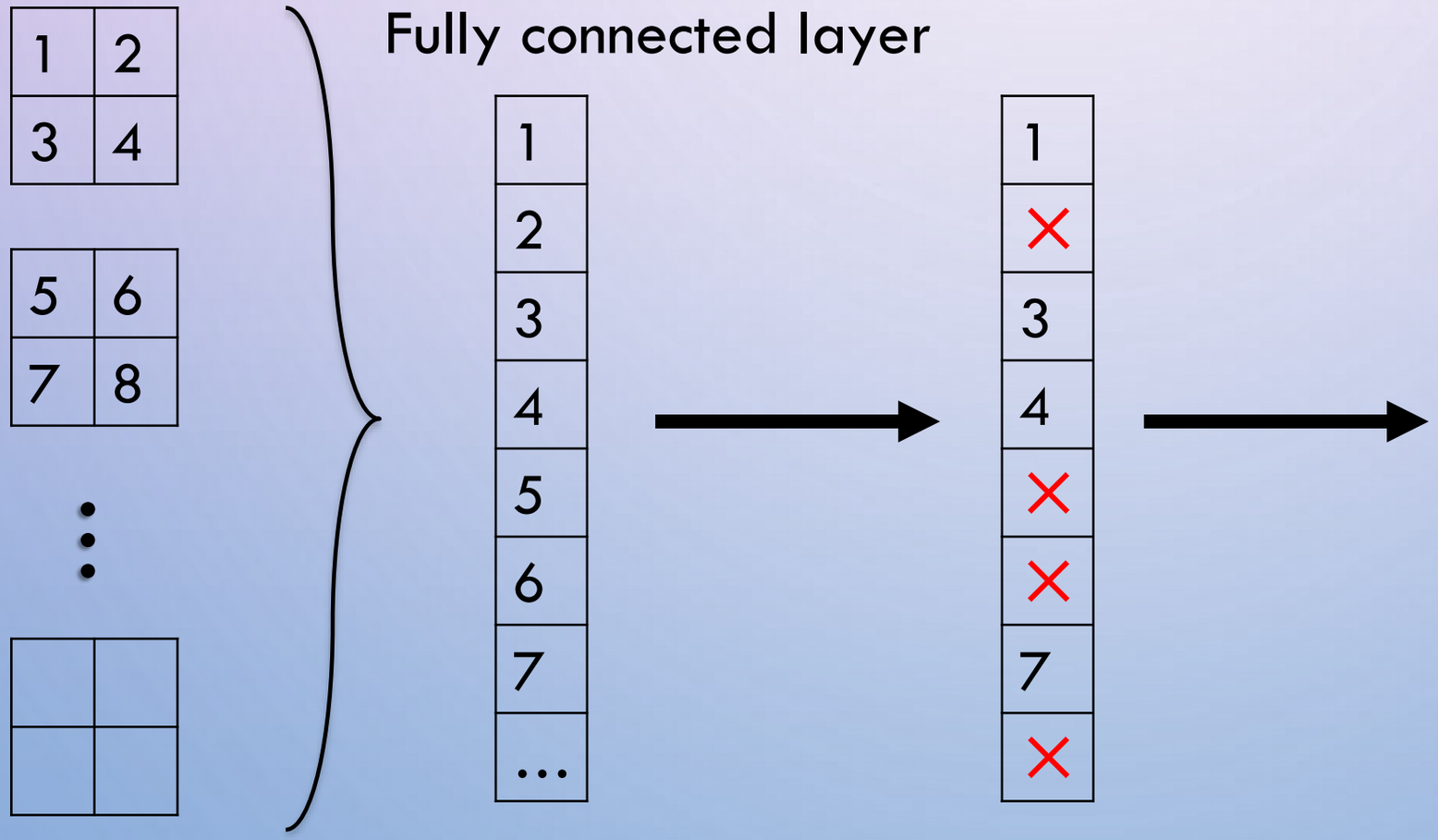
- **MAX POOLING**



* Reduce dimensionality and correct over-fitting

PROPOSED METHOD

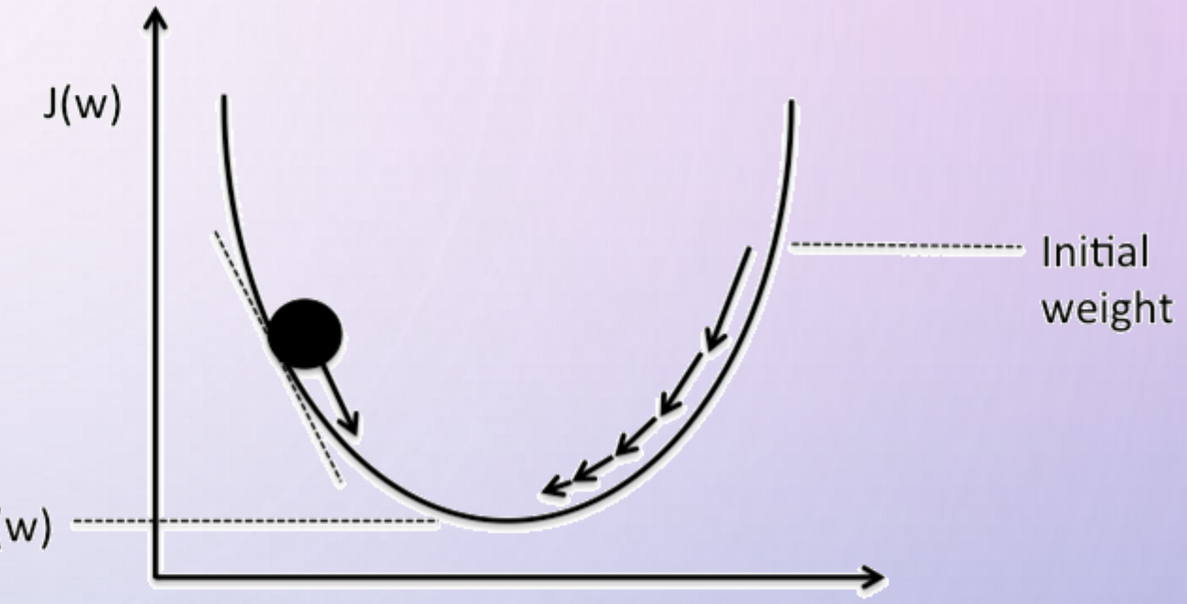
- **DROPOUT REGULARIZATION**



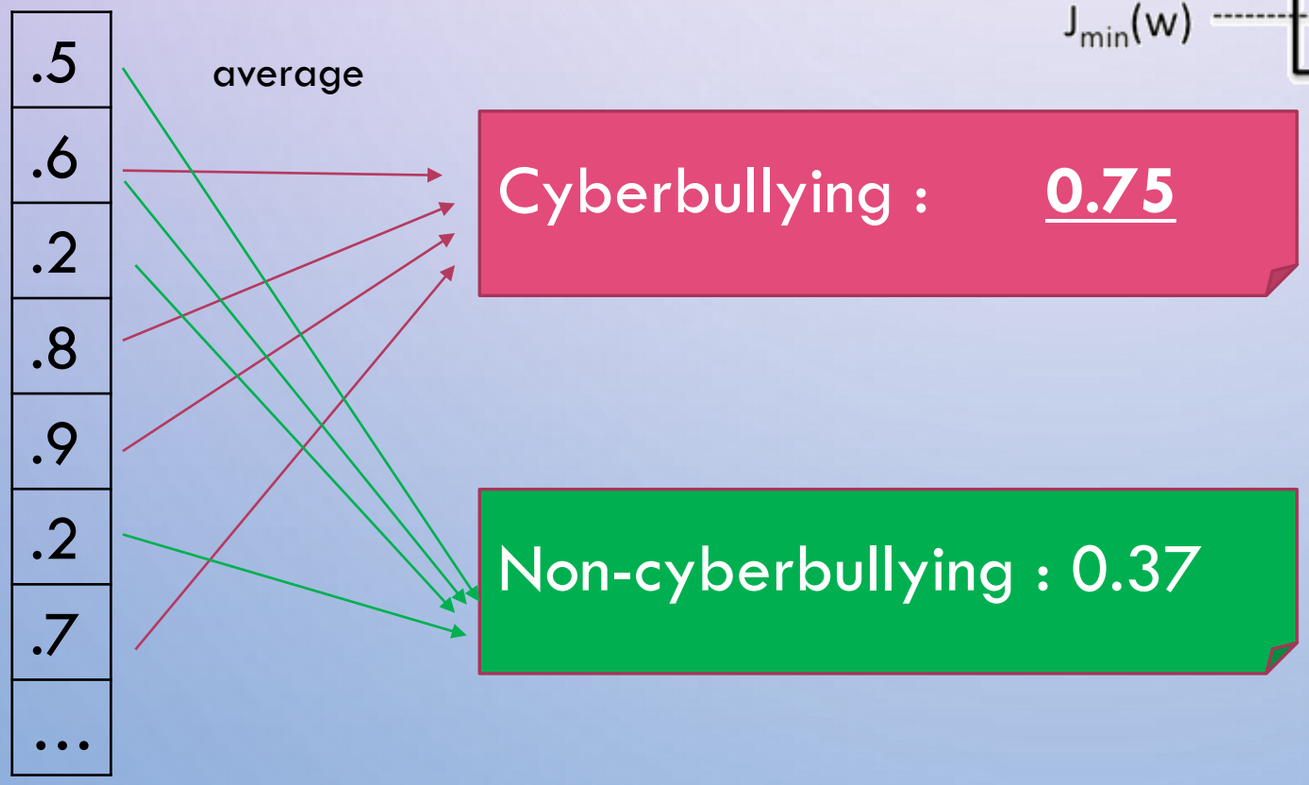
* Randomly delete some units during training to prevent co-adaptation of hidden units

PROPOSED METHOD

- **STOCHASTIC GRADIENT DESCENT**



Fully connected layer

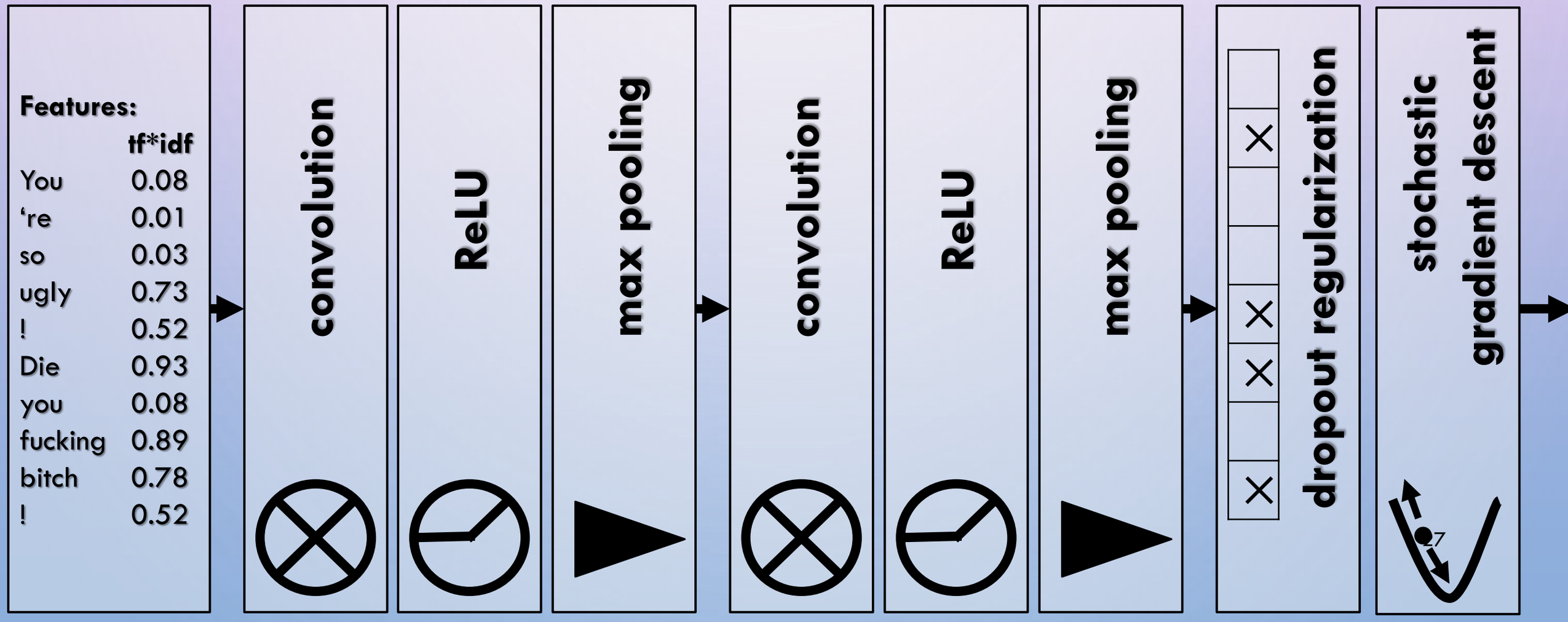


	correct	answer	Error
CB	1	.75	.25
N-CB	0	.37	.63
	Total error=		**0.88**

* Slide weight a little back and forth to minimize error

PROPOSED METHOD

- DEEP CONVOLUTIONAL NEURAL NETWORK



PROPOSED METHOD

- LANGAUGE COMBINATORICS
- DEEP CONVOLUTIONAL NEURAL NETWORK
- SHALLOW CNN
- SVM
  - LINEAR
  - POLYNOMIAL
  - RADIAL BASED FUNCTION
  - SIGMOID
- DECISION TREES (J48)
- RANDOM FORESTS
- KNN
- NAÏVE BAYES
- JRIP

[Ptaszynski et al., 2010]
[Sood et al., 2012]
[Dinakar et al., 2012]
[Sarna and Bhatia, 2015]
[Ptaszynski et al., 2015a,b]
…AND A FEW OTHERS

PROPOSED METHOD

- LANGAUGE COMBINATORICS → BRUTE FORCE
- DEEP CONVOLUTIONAL NEURAL NETWORK
- SHALLOW CNN → Neural Nets
- SVM
  - LINEAR
  - POLYNOMIAL
  - RADIAL BASED FUNCTION
  - SIGMOID

→ SVM

- DECISION TREES (J48)
- RANDOM FORESTS

→ trees

- KNN
- NAÏVE BAYES
- JRIP

→ lazy / rules

10-fold
X-validation

DATASET

- ACTUAL DATA COLLECTED BY INTERNET PATROL (ANNOTATED BY EXPERTS)
- FROM UNOFFICIAL SCHOOL FORUMS (BBS)
- PROVIDED BY HUMAN RIGHT CENTER IN JAPAN (MIE PREFECTURE)
- ACCORDING TO THE DEFINITION BY JAPANESE MINISTRY OF EDUCATION (MEXT)
- 1,490 HARMFUL AND 1,508 NON-HARMFUL ENTRIES

FEATURE SELECTION

<u>Example</u>: John McDonnald killed Mary Poppins!

- **TOKENIZATION:** JOHN MCDONNALD KILLED MARY POPPINS !
- **LEMMATIZATION:** JOHN MCDONNALD KILL MARY POPPINS !
- **PARTS OF SPEECH:** NOUN NOUN VERB NOUN NOUN EXCL.
- **TOKENS WITH POS:** JOHN_NOUN MCDONNALD_NOUN KILLED_VERB MARY_NOUN POPPINS_NOUN !_EXCL.
- **LEMMAS WITH POS:** JOHN_NOUN MCDONNALD_NOUN KILL_VERB MARY_NOUN POPPINS_NOUN !_EXCL.
- **TOKENS WITH NAMED ENTITY RECOGNITION:** JOHN [COMPANY] KILLED MARY POPPINS !
- **LEMMAS WITH NER:** JOHN [COMPANY] KILL MARY POPPINS !
- **CHUNKING:** JOHN_MCDONNALD_KILLED MARY_POPPINS_!
- **DEPENDENCY STRUCTURE:** 1-JOHN 1-MCDONNALD 2-KILLED 3-MARY 3-POPPINS 3-!
- **CHUNKING WITH NER:** JOHN_[COMPANY]_KILLED MARY_POPPINS_!
- **DEPENDENCY STRUCTURE WITH NAMED ENTITIES:** 1-JOHN 1-[COMPANY] 2-KILLED 3-MARY 3-POPPINS 3-!

RESULTS AND DISCUSSION

RESULTS AND DISCUSSION

- LAZY CLASSIFIERS :
  - GENERALLY POOR PERFORMANCE (F1 ~ 50% - 60%)
  - BETTER WITH NER (F1 ~ 70%)

RESULTS AND DISCUSSION

- LAZY CLASSIFIERS :
  - GENERALLY POOR PERFORMANCE (F1 ~ 50% - 60%)
  - BETTER WITH NER (F1 ~ 70%)
- TREE-BASED :
  - J48 – LOW
  - RANDOM FORREST – PRETTY GOOD (F1 ~ 80%), BUT TIME INEFFICIENT

RESULTS AND DISCUSSION

- LAZY CLASSIFIERS :
  - GENERALLY POOR PERFORMANCE (F1 ~ 50% - 60%)
  - BETTER WITH NER (F1 ~ 70%)
- TREE-BASED :
  - J48 – LOW
  - RANDOM FORREST – PRETTY GOOD (F1 ~ 80%), BUT TIME INEFFICIENT
- SVM
  - LINEAR – NICE (UP TO F1=82.5%)
  - OTHER – POOR
  - SUPER FAST TO TRAIN (BEST TIME TO PERFORMANCE RATIO)

RESULTS AND DISCUSSION

- BEST FEATURE SETS:
  - TOKENS + NER
  - LEMMAS + NER

RESULTS AND DISCUSSION

- BEST FEATURE SETS:
  - TOKENS + NER
  - LEMMAS + NER

** CYBERBULLYING IS OFTEN ABOUT REVEALING PRIVATE INFORMATION, NOT ONLY ABOUT SLANDERING **

RESULTS AND DISCUSSION

- BEST FEATURE SETS:
  - TOKENS + NER
  - LEMMAS + NER

** CYBERBULLYING IS OFTEN ABOUT REVEALING PRIVATE INFORMATION, NOT ONLY ABOUT SLANDERING **

- 2ND BEST METHOD (EXCEPT PROPOSED)
  - BRUTE FORCE (F1=80.3%)
  - EXCEPT ONE SVM CASE
    - LINEAR SVM ON LEMMAS (F1 = 82.5%)

RESULTS AND DISCUSSION

- BEST METHOD
  - CNN WITH 2HIDDEN LAYERS (PROPOSED)
  - F1=93.5%
  - NER ALWAYS HELPED

RESULTS AND DISCUSSION

- BEST METHOD
  - CNN WITH 2HIDDEN LAYERS (PROPOSED)
  - F1=93.5%
  - NER ALWAYS HELPED

…WHY?

RESULTS AND DISCUSSION

- GREATEST PROBLEM WITH NEURAL NETS
  (AND ALSO MANY OTHER MACHINE LEARNING METHODS)

Q

RESULTS AND DISCUSSION

- GREATEST PROBLEM WITH NEURAL NETS
  (AND ALSO MANY OTHER MACHINE LEARNING METHODS)

- INTERPRETABILITY
  - WHY RESULTS WERE AS GOOD?
  - WHAT EXACTLY WAS SO GOOD ABOUT IT?
  - WHAT INFLUENCED THE RESULTS?

RESULTS AND DISCUSSION

- GREATEST PROBLEM WITH NEURAL NETS
  (AND ALSO MANY OTHER MACHINE LEARNING METHODS)

- INTERPRETABILITY
  - WHY RESULTS WERE AS GOOD?
  - WHAT EXACTLY WAS SO GOOD ABOUT IT?
  - WHAT INFLUENCED THE RESULTS?

IF YOU DON'T KNOW THE CAUSE YOU CAN ALWAYS LOOK FOR CORRELATION

RESULTS AND DISCUSSION

1. NAMED ENTITY RECOGNITION USUALLY HELPED ESPECIALLY WITH CNN
2. …?

→GENERAL LOOK AT THE DATA

→WHAT IS DIFFERENT ABOUT THE DATA?

RESULTS AND DISCUSSION

- LEXICAL DENSITY [URE, 1971]

ALL UNIQUE WORDS IN CORPUS / ALL WORDS IN CORPUS

[Ure, 1971] J. Ure. Lexical density and register differentiation. Applications of Linguistics, pages 443–452, 1971.

RESULTS AND DISCUSSION

- LEXICAL DENSITY [URE, 1971]

ALL UNIQUE WORDS IN CORPUS / ALL WORDS IN CORPUS

NOT JUST WORDS:

TOKENS, LEMMAS, POS, TOKEN-POS, TOKEN-NER, LEMMA-POS, LEMMA-NER, CHUNKS, CHUNKS-NER, DEPENDENCY, DEP-NER,

[Ure, 1971] J. Ure. Lexical density and register differentiation. Applications of Linguistics, pages 443–452, 1971.

RESULTS AND DISCUSSION

- ~~LEXICAL DENSITY [URE, 1971]~~ → **FEATURE DENSITY**

ALL UNIQUE WORDS IN CORPUS / ALL WORDS IN CORPUS

<u>NOT JUST WORDS:</u>

TOKENS, LEMMAS, POS, TOKEN-POS, TOKEN-NER, LEMMA-POS, LEMMA-NER, CHUNKS, CHUNKS-NER, DEPENDENCY, DEP-NER,

[Ure, 1971] J. Ure. Lexical density and register differentiation. Applications of Linguistics, pages 443–452, 1971.

FEATURE DENSITY

- CALCULATE FD FOR ALL DATASETS
- CHECK CORRELATION BETWEEN FD AND CLASSIFIER RESULTS

FEATURE DENSITY

Classifier	ρ value	2-sided p-value
CNN-2L	0.638	*p=0.035
SVM-pol	-0.431	p=0.185
SVM-sig	-0.534	p=0.091
SPEC(BEP)	-0.550	p=0.133
RF	-0.560	p=0.073
SVM-lin	-0.564	p=0.076
SPEC(F1)	-0.636	p=0.066
SVM-rad	-0.639	*p=0.034
CNN-1L	-0.709	*p=0.019
JRip	-0.729	*p=0.011
NB	-0.736	*p=0.013
J48	-0.791	**p=0.006
kNN	-0.809	**p=0.004

FEATURE DENSITY

Classifier	ρ value	2-sided p-value
CNN-2L	0.638	*p=0.035
SVM-pol	-0.431	p=0.185
SVM-sig	-0.534	p=0.091
SPEC(BEP)	-0.550	p=0.133
RF	-0.560	p=0.073
SVM-lin	-0.564	p=0.076
SPEC(F1)	-0.636	p=0.066
SVM-rad	-0.639	*p=0.034
CNN-1L	-0.709	*p=0.019
JRip	-0.729	*p=0.011
NB	-0.736	*p=0.013
J48	-0.791	**p=0.006
kNN	-0.809	**p=0.004

DENSIER DATA KILLS CLASSIFIER

POORLY PERFORMING CLASSIFIERS: NEGATIVE STRONG CORELATION WITH FD

FEATURE DENSITY

Classifier	ρ value	2-sided p-value
CNN-2L	0.638	*p=0.035
SVM-pol	-0.431	p=0.185
SVM-sig	-0.534	p=0.091
SPEC(BEP)	-0.550	p=0.133
RF	-0.560	p=0.073
SVM-lin	-0.564	p=0.076
SPEC(F1)	-0.636	p=0.066
SVM-rad	-0.639	*p=0.034
CNN-1L	-0.709	*p=0.019
JRip	-0.729	*p=0.011
NB	-0.736	*p=0.013
J48	-0.791	**p=0.006
kNN	-0.809	**p=0.004

PROBABLY* **DENSIER DATA HARMS CLASSIFIER**

(*low significance)

MODERATELY PERFORMING CLASSIFIERS:
NEGATIVE MEDIUM CORELATION FITH FD

FEATURE DENSITY

Classifier	ρ value	2-sided p-value
CNN-2L	0.638	*p=0.035
SVM-pol	-0.431	p=0.185
SVM-sig	-0.534	p=0.091
SPEC(BEP)	-0.550	p=0.133
RF	-0.560	p=0.073
SVM-lin	-0.564	p=0.076
SPEC(F1)	-0.636	p=0.066
SVM-rad	-0.639	*p=0.034
CNN-1L	-0.709	*p=0.019
JRip	-0.729	*p=0.011
NB	-0.736	*p=0.013
J48	-0.791	**p=0.006
kNN	-0.809	**p=0.004

→ CNN – BEST PERFORMANCE
STRONG POSITIVE
CORRELATION WITH FD

DENSIER DATA = BETTER RESULTS

FEATURE DENSITY

Classifier	ρ value	2-sided p-value
CNN-2L	0.638	*p=0.035
SVM-pol	-0.431	p=0.185
SVM-sig	-0.534	p=0.091
SPEC(BEP)	-0.550	p=0.133
RF	-0.560	p=0.073
SVM-lin	-0.564	p=0.076
SPEC(F1)	-0.636	p=0.066
SVM-rad	-0.639	*p=0.034
CNN-1L	-0.709	*p=0.019
JRip	-0.729	*p=0.011
NB	-0.736	*p=0.013
J48	-0.791	**p=0.006
kNN	-0.809	**p=0.004

→ CNN – BEST PERFORMANCE
STRONG POSITIVE
CORRELATION WITH FD

FUTURE:
LET'S CHECK
EVEN
DENSIER DATA

CONCLUSIONS

- PROBLEM: CYBERBULLYING DETECTION
- MANY FEATURE SETS
- MANY CLASSIFIERS
- PROPOSED DEEP CNN SOLUTION
- NAMED ENTITY ANNOTATION USUALLY HELPED DETECT CYBERBULLYING
- FEATURE DENSITY POSITIVELY CORELATED WITH CNN RESULTS
  - WILL CHECK EVEN DENSIER FEATURE SETS
  - WILL CHECK FOR OTHER TASKS : SENTIMENT, DECEPTION, SARCASM, ETC.

THANK YOU FOR YOUR KIND ATTENTION!

MICHAL PTASZYNSKI

PTASZYNSKI@IEEE.ORG