INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

NLPIR 2024
8th International Conference on Natural Language Processing and Information Retrieval

JULIA HEINE, KRISTINA SCHAAFF AND TIM SCHLIPPE

INVESTIGATING TEXT AND IMAGE FEATURES FOR THE DETECTION OF CYBERBULLYING

Okayama, Japan
December 14, 2024

AGENDA

1

INTRODUCTION

MOTIVATION

~50% OF ALL TEENAGERS HAVE EXPERIENCE WITH CYBERBULLYING

Text Source: Vogels (2022).

MOTIVATION

~50% OF ALL TEENAGERS HAVE EXPERIENCE WITH CYBERBULLYING

➔ trouble sleeping, difficulty concentrating, or feelings of anxiety

Text Source: Vogels (2022).

MOTIVATION

~50% OF ALL TEENAGERS HAVE EXPERIENCE WITH CYBERBULLYING

➔ trouble sleeping, difficulty concentrating, or feelings of anxiety

CYBERBULLYING CAN BE MULTIMODAL!

Text Source: Vogels (2022).

MOTIVATION

~50% OF ALL TEENAGERS HAVE EXPERIENCE WITH CYBERBULLYING

➔ trouble sleeping, difficulty concentrating, or feelings of anxiety

CYBERBULLYING CAN BE MULTIMODAL!

➔ text, images, videos

Text Source: Vogels (2022).

MOTIVATION

CHALLENGE: TOO MUCH CONTENT FOR MANUAL DETECTION

Text Source: Vogels (2022).

MOTIVATION

CHALLENGE: TOO MUCH CONTENT FOR MANUAL DETECTION

➔ volume of content shared online makes it impossible to manually monitor and address every violation

Text Source: Vogels (2022).

MOTIVATION

CHALLENGE: TOO MUCH CONTENT FOR MANUAL DETECTION

➔ volume of content shared online makes it impossible to manually monitor and address every violation

SOLUTION: AI TO DETECT CYBERBULLYING AUTOMATICALLY

Text Source: Vogels (2022).

MOTIVATION

CHALLENGE: TOO MUCH CONTENT FOR MANUAL DETECTION

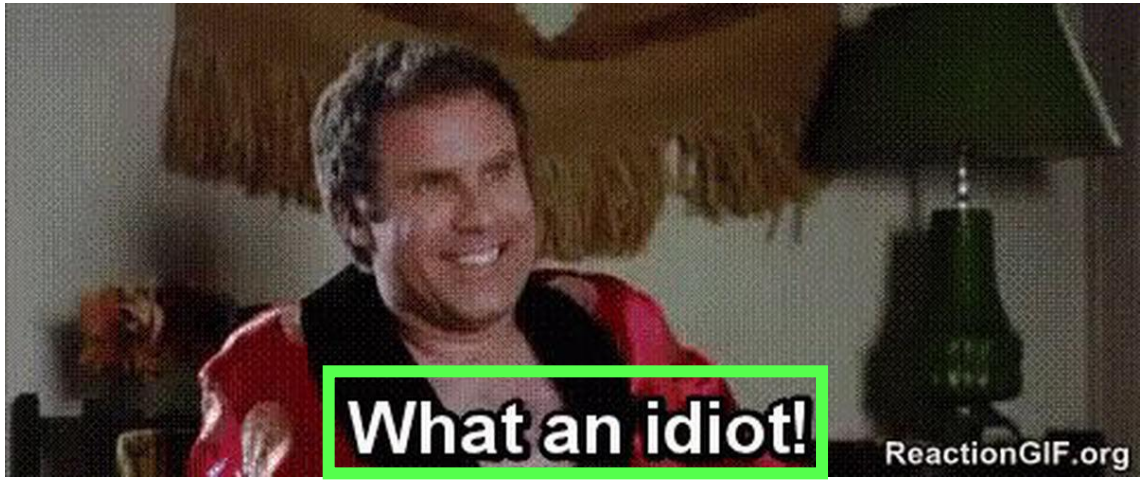
➔ volume of content shared online makes it impossible to manually monitor and address every violation

SOLUTION: AI TO DETECT CYBERBULLYING AUTOMATICALLY

➔ automatically detect cyberbullying and address it more effectively and at scale

Text Source: Vogels (2022).

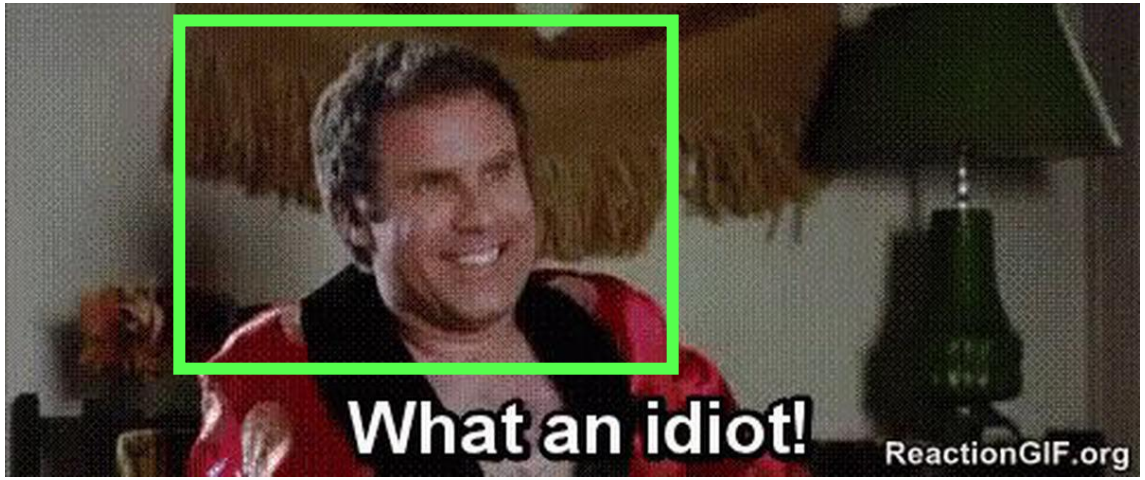
MOTIVATION: RESEARCH QUESTIONS



1. WHAT TEXT FEATURES HELP TO IMPROVE CYBERBULLYING DETECTION?

Text Source: Vogels (2022).

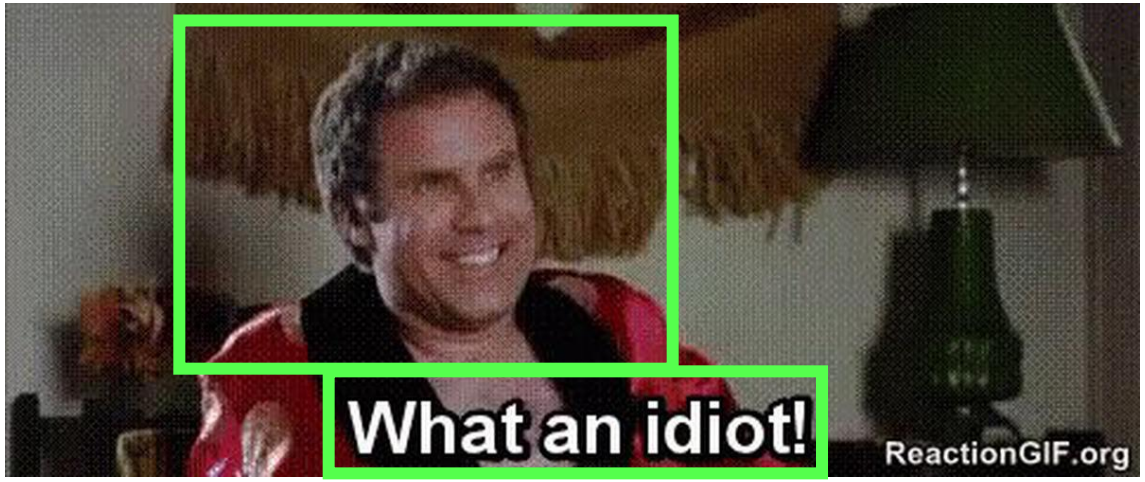
MOTIVATION: RESEARCH QUESTIONS



2. WHAT IMAGE FEATURES HELP TO IMPROVE CYBERBULLYING DETECTION?

Text Source: Vogels (2022).

MOTIVATION: RESEARCH QUESTIONS



3. WHAT IS THE PERFORMANCE OF THE COMBINATION OF TEXT AND IMAGE FEATURES FOR CYBERBULLYING DETECTION?

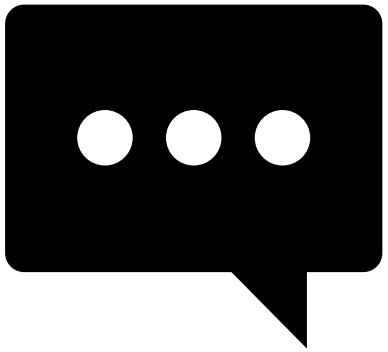
Text Source: Vogels (2022).

2

RELATED WORK

RELATED WORK

Text features are more investigated

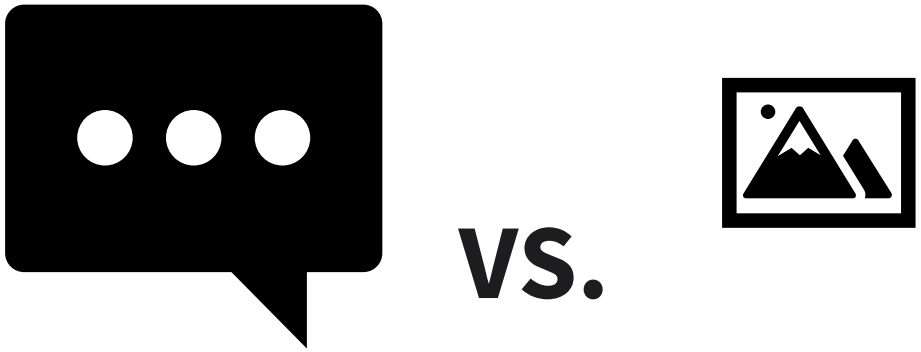


VS.

e.g., Batani et al. (2022), Islam et al. (2020), Raj et al. (2021), Alkomah et al. (2022), Atoum (2020), Van Bruwaene et al. (2020)

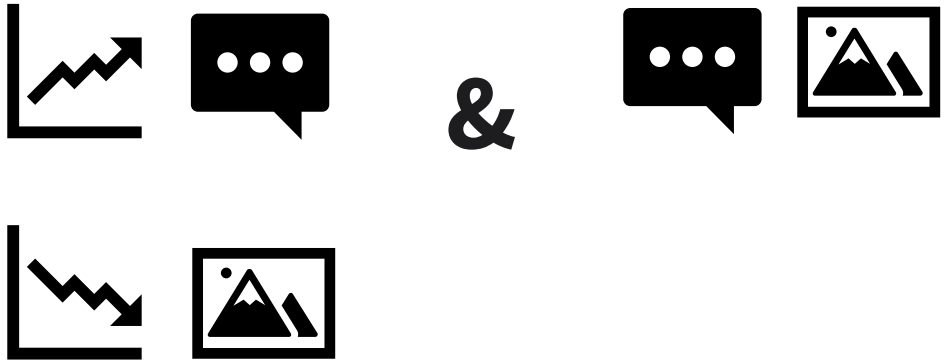
RELATED WORK

Text features are more investigated



e.g., Batani et al. (2022), Islam et al. (2020), Raj et al. (2021), Alkomah et al. (2022), Atoum (2020), Van Bruwaene et al. (2020)

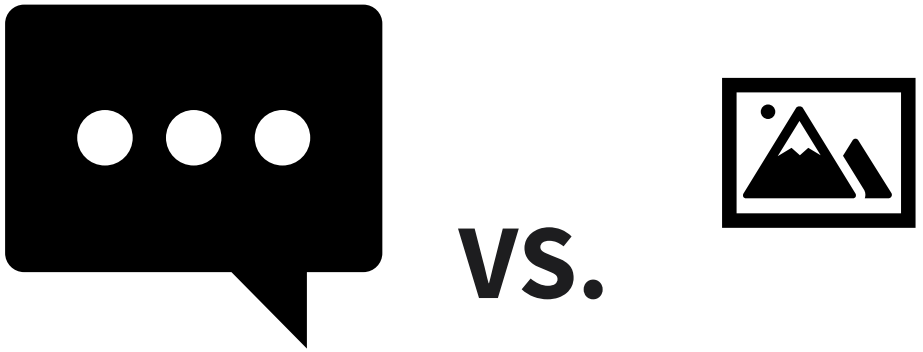
Image Models are not that accurate



e.g., Hosseinmardi et al. (2016), Gomez et al. (2019), Botelho et al. (2021), Sahu et al. (2021), Paul et al. (2022)

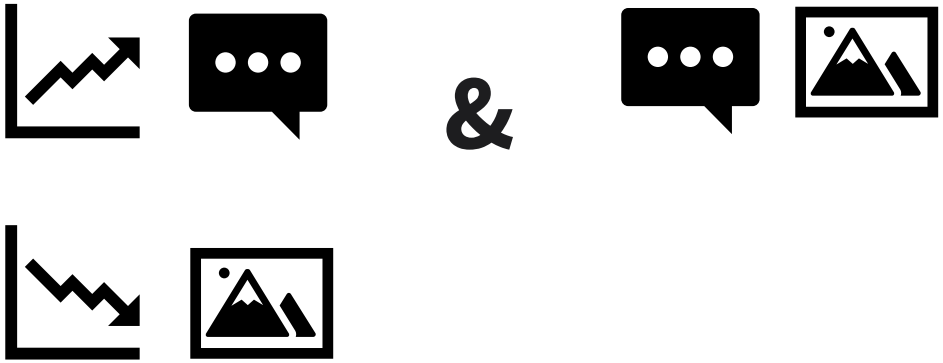
RELATED WORK

Text features are more investigated



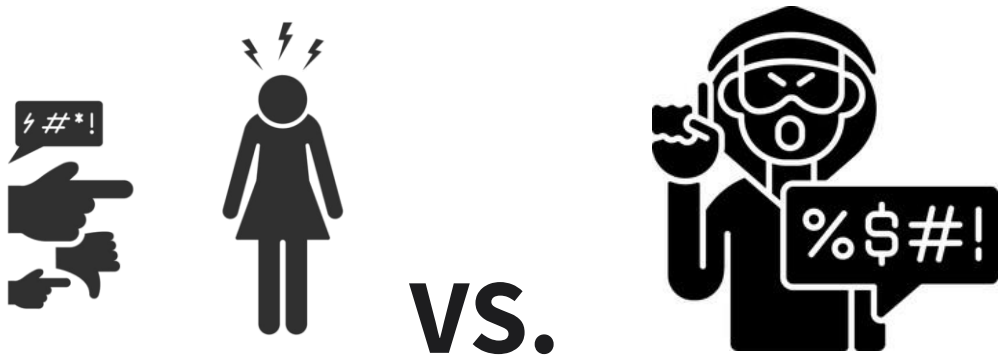
e.g., Batani et al. (2022), Islam et al. (2020), Raj et al. (2021), Alkomah et al. (2022), Atoum (2020), Van Bruwaene et al. (2020)

Image Models are not that accurate



e.g., Hosseinmardi et al. (2016), Gomez et al. (2019), Botelho et al. (2021), Sahu et al. (2021), Paul et al. (2022)

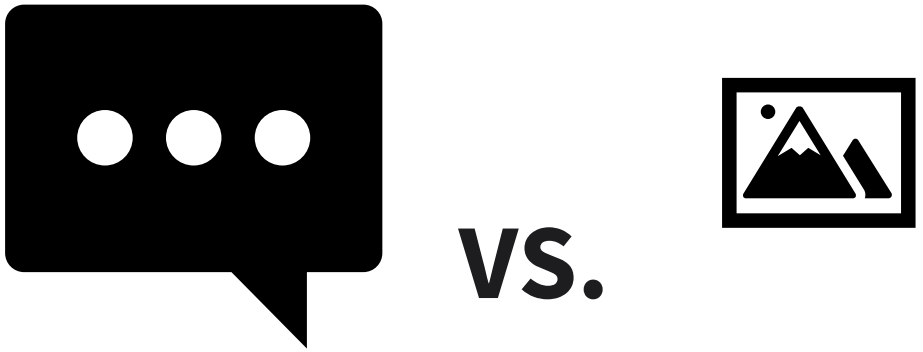
Cyberbullying = Hate Speech



e.g., Batani et al. (2022)

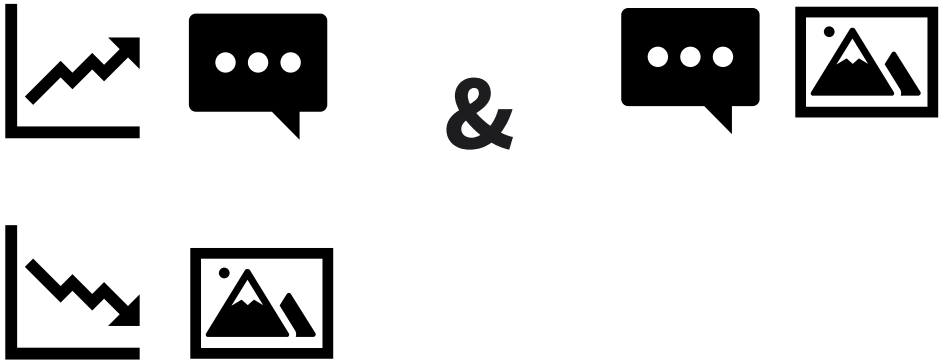
RELATED WORK

Text features are more investigated



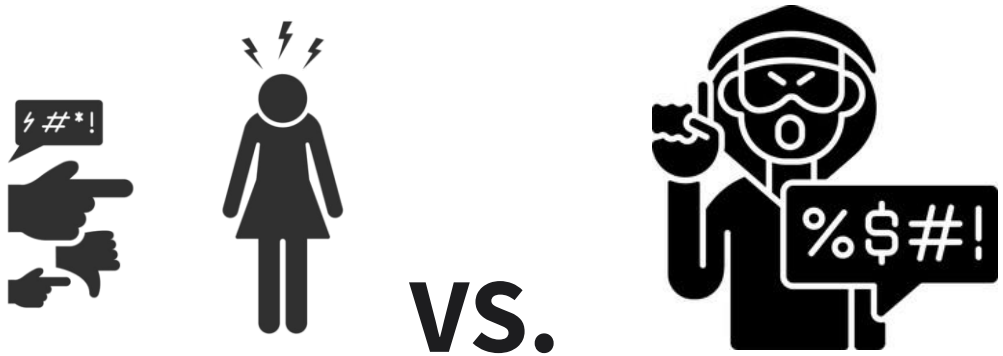
e.g., Batani et al. (2022), Islam et al. (2020), Raj et al. (2021), Alkomah et al. (2022), Atoum (2020), Van Bruwaene et al. (2020)

Image Models are not that accurate



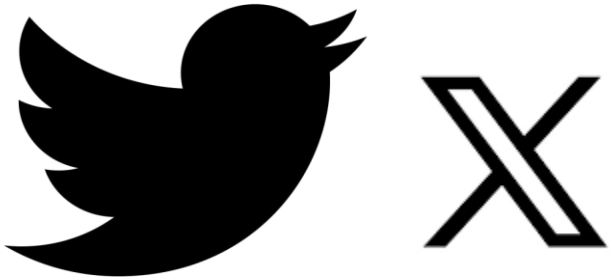
e.g., Hosseinmardi et al. (2016), Gomez et al. (2019), Botelho et al. (2021), Sahu et al. (2021), Paul et al. (2022)

Cyberbullying = Hate Speech



e.g., Batani et al. (2022)

Most datasets are based on Twitter



e.g., Gomez et al. (2019), Mahlangu & Tu (2019), Yadav et al. (2020), Mahat (2021), Roy et al. (2020), Bharti et al. (2021)

EXPERIMENTAL SETUP

EXPERIMENTAL SETUP: GOAL & APPROACH

GOAL

Develop a simple and structured approach to answer research questions

EXPERIMENTAL SETUP: GOAL & APPROACH

GOAL

Develop a simple and structured approach to answer research questions

APPROACH

Multimodel dataset

➔ MMHS150K dataset (Gomez et al., 2019)

EXPERIMENTAL SETUP: GOAL & APPROACH

GOAL

Develop a simple and structured approach to answer research questions

APPROACH

Multimodel dataset
➔ MMHS150K dataset (Gomez et al., 2019)

Binary Classification Problem
➔ Convert 6-class dataset into binary labels

EXPERIMENTAL SETUP: GOAL & APPROACH

GOAL

Develop a simple and structured approach to answer research questions

APPROACH

Multimodel dataset
➔ MMHS150K dataset (Gomez et al., 2019)

Binary Classification Problem
➔ Convert 6-class dataset into binary labels

Balanced Dataset
➔ Ensure equal distribution of *bullying* and *non-bullying* samples across train, validation, and test sets

EXPERIMENTAL SETUP: GOAL & APPROACH

GOAL

Develop a simple and structured approach to answer research questions

APPROACH

Multimodel dataset
➔ MMHS150K dataset (Gomez et al., 2019)

Binary Classification Problem
➔ Convert 6-class dataset into binary labels

Balanced Dataset
➔ Ensure equal distribution of *bullying* and *non-bullying* samples across train, validation, and test sets

Simple Classifier
➔ Multi-Layer Perceptron (MLP)

EXPERIMENTAL SETUP: FEATURE DESIGN & TESTING

FEATURE DESIGN

- Combine features from Hate Speech and Cyberbullying literature
- Create new features

EXPERIMENTAL SETUP: FEATURE DESIGN & TESTING

FEATURE DESIGN

- Combine features from Hate Speech and Cyberbullying literature
- Create new features

TESTING MODELS

Single-Feature Models

➔ Evaluate performance of individual features

EXPERIMENTAL SETUP: FEATURE DESIGN & TESTING

FEATURE DESIGN

- Combine features from Hate Speech and Cyberbullying literature
- Create new features

TESTING MODELS

Single-Feature Models
➔ Evaluate performance of individual features

Multi-Feature Models
➔ Test combinations of features to improve performance

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: TEXT FEATURES

Text Vector Features

Sentence Vector

Error-Based Features

Semantic Features

Objectivity vs. Subjectivity
Sentiment
Sarcasm
Toxicity

List Lookup Features

Negative words
Special characters
Exclamation/question marks
Uppercase letters

Emoji Features

Negative

Neutral

Positive

AI Feedback Features

EXPERIMENTAL SETUP: IMAGE FEATURES

Image Vector Features

Image Vector

Image Text Features

Sentence Vector

Image Description Features

1st best description
1st best description + Probability
5 best descriptions

EXPERIMENTAL SETUP: IMAGE FEATURES

Image Vector Features

Image Vector

Image Text Features

Sentence Vector

Image Description Features

1st best description
1st best description + Probability
5 best descriptions

EXPERIMENTAL SETUP: IMAGE FEATURES

Image Vector Features

Image Vector

Image Text Features

Sentence Vector

Image Description Features

1st best description
1st best description + Probability
5 best descriptions

EXPERIMENTAL SETUP: IMAGE FEATURES

Image Vector Features

Image Vector

Image Text Features

Sentence Vector

Image Description Features

1st best description
1st best description + Probability
5 best descriptions

EXPERIMENTS AND RESULTS

RESULTS

Feature Category	F1
Text Vector Features	66.25%
Semantic Features	56.24%
- $sentiment_{score}$	39.92%
- $subjectivity_{score}$*	43.43%
- $sarcasm_{score}$*	53.21%
- $toxicity_{score}$*	54.41%
List Lookup Features	51.90%
- $capitalizedChars_{rel}$	15.17%
- $negWords_{rel}$	42.10%
- $specialChars_{rel}$*	66.30%
Emoji Features	56.53%
- $negEmoji_{score}$	66.29%
- $neuEmoji_{score}$	**66.39%**
- $posEmoji_{score}$	65.54%
*Error-Based Features**	56.76%
*AI Feedback Features**	64.45%
$\text{Text Features}_{all}$	65.54%
$\text{Text Features}_{F1>50\%}$	66.08%

*** NEW FEATURES**

RESULTS

Feature Category	F1
Text Vector Features	66.25%
Semantic Features	56.24%
- $sentiment_{score}$	39.92%
- $subjectivity_{score}$*	43.43%
- $sarcasm_{score}$*	53.21%
- $toxicity_{score}$*	54.41%
List Lookup Features	51.90%
- $capitalizedChars_{rel}$	15.17%
- $negWords_{rel}$	42.10%
- $specialChars_{rel}$*	66.30%
Emoji Features	56.53%
- $negEmoji_{score}$	66.29%
- $neuEmoji_{score}$	**66.39%**
- $posEmoji_{score}$	65.54%
*Error-Based Features**	56.76%
*AI Feedback Features**	64.45%
$\text{Text Features}_{all}$	65.54%
$\text{Text Features}_{F1>50\%}$	66.08%

> 66% F1

*** NEW FEATURES**

RESULTS

Feature Category	F1
Text Vector Features	66.25%
Semantic Features	56.24%
- $sentiment_{score}$	39.92%
- $subjectivity_{score}$*	43.43%
- $sarcasm_{score}$*	53.21%
- $toxicity_{score}$*	54.41%
List Lookup Features	51.90%
- $capitalizedChars_{rel}$	15.17%
- $negWords_{rel}$	42.10%
- $specialChars_{rel}$*	66.30%
Emoji Features	56.53%
- $negEmoji_{score}$	66.29%
- $neuEmoji_{score}$	**66.39%**
- $posEmoji_{score}$	65.54%
*Error-Based Features**	56.76%
*AI Feedback Features**	64.45%
$\text{Text Features}_{all}$	65.54%
$\text{Text Features}_{F1>50\%}$	66.08%

> 60% F1

* NEW FEATURES

Feature Category	F1
Image Vector Features	**65.27%**
Image Text Features	31.57%
Image Description Features	60.92%
- $firstBestDescription$	25.11%
- $firstBestDescriptionProb$*	62.37%
- $fiveBestDescription$*	28.72%
$\text{Image Features}_{all}$	30.39%
$\text{Image Features}_{F1>50\%}$	56.10%

RESULTS

Feature Category	F1
Text Vector Features	66.25%
Semantic Features	56.24%
- $sentiment_{score}$	39.92%
- $subjectivity_{score}$*	43.43%
- $sarcasm_{score}$*	53.21%
- $toxicity_{score}$*	54.41%
List Lookup Features	51.90%
- $capitalizedChars_{rel}$	15.17%
- $negWords_{rel}$	42.10%
- $specialChars_{rel}$*	66.30%
Emoji Features	56.53%
- $negEmoji_{score}$	66.29%
- $neuEmoji_{score}$	**66.39%**
- $posEmoji_{score}$	65.54%
*Error-Based Features**	56.76%
*AI Feedback Features**	64.45%
Text $Features_{all}$	65.54%
Text $Features_{F1>50\%}$	66.08%

Feature Category	F1
Image Vector Features	**65.27%**
Image Text Features	31.57%
Image Description Features	60.92%
- $firstBestDescription$	25.11%
- $firstBestDescriptionProb$*	62.37%
- $fiveBestDescription$*	28.72%
Image $Features_{all}$	30.39%
Image $Features_{F1>50\%}$	56.10%

*** NEW FEATURES**

Feature Category	F1
Text+Image $Features_{all}$	64.42%
Text+Image $Features_{F1>50\%}$	65.37%

5

CONCLUSION AND FUTURE WORK

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.
- Best image-based system achieved 65.27% F1 using image vector features.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.
- Best image-based system achieved 65.27% F1 using image vector features.
- Our innovative features provided valuable insights but limited improvement due to task difficulty.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.
- Best image-based system achieved 65.27% F1 using image vector features.
- Our innovative features provided valuable insights but limited improvement due to task difficulty.

Future Work

- Evaluate features with different classifiers.
- Test features on other datasets.
- Extend features to other domains.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.
- Best image-based system achieved 65.27% F1 using image vector features.
- Our innovative features provided valuable insights but limited improvement due to task difficulty.

Future Work

- Evaluate features with different classifiers.
- Test features on other datasets.
- Extend features to other domains.

CONCLUSION AND FUTURE WORK

Conclusion

- We have explored features for distinguishing between bullying and non-bullying messages with images.
- Best text-based systems achieved over 66% F1 on the MMSH150K test set.
- Successful text features: Text vector features, list lookup features, emoji features plus the combination of these features.
- Best image-based system achieved 65.27% F1 using image vector features.
- Our innovative features provided valuable insights but limited improvement due to task difficulty.

Future Work

- Evaluate features with different classifiers.
- Test features on other datasets.
- Extend features to other domains.

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

THANK YOU

Tim Schlippe
tim.schlippe@iu.org