# How Can We **Detect Toxicity for Korean**?:
## Toxic Comments Classification for Korean Movie Comments

20150497 Wi Heeju

# Motivation

# Why AI Needed?

1. Too many comments updated per second,
   **so human cannot check all** the comments.

2. We have to detect the toxicity of the comments
   **before human see it**, to prevent the harassment.

3. Everyone have **different baseline** of considering the toxicity of the comment,
   using classifier can be a solution to remove controversial problem.

4. Toxicity depends on **the context** of the comment,
   so we have to consider not only the word itself but also the context.

Introduction

# For English..

**Jigsaw (Google)**

built **Perspective API**

**to score the toxicity of the comment**

# Toxic Comments Classification Challenge


**Dataset for the Challenge**

•comments from Wikipedia's talk page edits

•classified in 6 categories(toxic, severe toxic, obscene, threat, insult, identity hate)

•160K comments for training set

# Even Perspective..

Jigsaw, preparing for Korean service

## 구글 자회사, 한국어 악플 차단도 준비

**제라드 코엔 직소 대표 기자회견**

**2017.09.13**

직소 대표인 제라드 코엔(Jared Cohen)은 한국어에 대한 충분한 데이터가 모이면 한국어 서비스도 할 의향이 있음을 밝혔다.

코엔은 12일 대전 컨벤션센터에서 가진 기자회견에서 "현재 스페인어는 준비 중이고 한국어에 대한 충분한 데이터가 모이면 하겠다"고 말했다.

https://www.sciencetimes.co.kr/?news=%EA%B5%AC%EA%B8%80-%EC%9E%90%ED%9A%8C%EC%82%AC-%ED%95%9C%EA%B5%AD%EC%96%B4-%EC%95%85%ED%94%8C-%EC%B0%A8%EB%8B%A8%EB%8F%84-%EC%A4%80%EB%B9%84

# For Korean..?

**No Public** Korean Toxic comments **Data**

**No Public** Korean Toxic Comments **Classifier**

**Surprisingly little work related to** Korean hate speech

# Research Statement

Making Korean Toxic Classifier
using Korean Movie comments

# Prior Research

Prior Research

# Definitions

**Hate speech**
text to **foster hate** against specific individuals/organizations, by causing a sounding board effect, which may critically damage the targets of the hate campaign, by using both psychological and physical violence.

**Cyberbullying**
an **aggressive, intentional** act or behavior that is carried out by a group or an individual, using electronic forms of contact, repeatedly and over time **against a victim** who cannot easily defend him or herself.
Ex) cyberstalking, trolling

**Online harassment**
All of harassment can be done online.
Ex) Spam mail, instant messages, website entries

**Toxic comments**

Looking at toxicity of online comments (Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale.) related research includes the investigation of hate speech (Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In WWW.; Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics; Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. Automated hate speech detection and the problem of offensive language), online harassment (Golbeck et al., 2017. A large labeled corpus for online harassment research.), abusive language (Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words?; Ji Ho Park and Pascale Fung. 2017. One-step and two- step classification for abusive language detection on twitter.), cyberbullying (Dadvar et al., 2013; Dinakar et al., 2012; Hee et al., 2015; Zhong et al., 2016) and offensive language (Chen et al., 2012; Xiang et al., 2012).

Prior Research

# 6 categories of Toxicity

| label | example |
|---|---|
| severe_toxic (심각) | "good job for sucking dick<br>dick trophy i dont have to do shit u say . and ur the worlds best dick sucker" (+obscene) |
| obscene (외설적인) | "DUDE!!! CALM THE FUCK DOWN!!!" |
| threat (협박) | "you just wait, your death is near" |
| insult (모욕) | "Go Fuck Yourself<br>Get a job, you hippie shitbag." (+obscene) |
| identity_hate (혐오) | "Hurry and ban me or protect this page you homo bitches" (+obscene, insult) |

Looking at toxicity of online comments (Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale.) related research includes the investigation of hate speech (Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In WWW.; Pete Burnap and Matthew L. Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics; Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. Automated hate speech detection and the problem of offensive language), online harassment (Golbeck et al., 2017. A large labeled corpus for online harassment research.), abusive language (Yashar Mehdad and Joel R. Tetreault. 2016. Do characters abuse more than words?; Ji Ho Park and Pascale Fung. 2017. One-step and two- step classification for abusive language detection on twitter.), cyberbullying (Dadvar et al., 2013; Dinakar et al., 2012; Hee et al., 2015; Zhong et al., 2016) and offensive language (Chen et al., 2012; Xiang et al., 2012).

# Prior Research
# On Korean Toxicity comments

Prior Research

# Difference between Korean and English toxic comments (2019)

**1. Different using rate of** 1st person **pronoun** and 3rd person pronoun.

English toxic comments : use of 3rd person pronoun > use of 1st person pronoun

Korean toxic comments : use of 3rd person pronoun = use of 1st person pronoun

**2. In specific**(social, religion, sadness, depression) **topic, having different rate of toxic comments**

English toxic comments > English civil comments

Korean toxic comments < Korean civil comments

**3. Having different frequency of specific vocabularies**(body, bio, ingest)

Korean toxic comments > frequency of English toxic comments

Young-il Kim, Youngjun Kim, Youngjin Kim and Kyungil Kim, "The Characteristics of Malicious Comments : Comparisons of the Internet News Comments in Korean and English," JOURNAL OF THE KOREA CONTENTS ASSOCIATION, Vol. 19, No. 1, pp. 548~558, Jan, 2019.

Prior Research

# Difference between Korean and English toxic comments (2019)

**1. Different using rate of** 1$^{st}$ person **pronoun** and 3$^{rd}$ person pronoun.

English toxic comments : use of 3rd person pronoun > use of 1st person pronoun

Korean toxic comments : use of 3rd person pronoun = use of 1st person pronoun

**2. In specific**(social, religion, sadness, depression) **topic, having different rate of toxic comments**

English toxic comments > English civil comments

Korean toxic comments < Korean civil comments

**3. Having different frequency of specific vocabularies**(body, bio, ingest)

Korean toxic comments > frequency of English toxic comments

**Translate Korean into English and using English toxic classifier wouldn't work!**

Young-il Kim, Youngjun Kim, Youngjin Kim and Kyungil Kim, "The Characteristics of Malicious Comments : Comparisons of the Internet News Comments in Korean and English," JOURNAL OF THE KOREA CONTENTS ASSOCIATION, Vol. 19, No. 1, pp. 548~558, Jan, 2019.

# Morpheme + NN Model (2019.07)

**표 5.** 1:1 데이터셋 : 분류 모델의 성능 비교

**Table 5.** 1:1 Dataset: Performance Comparison of Classification Mo

| | Accuracy | | | | Precision | | | |
|---|---|---|---|---|---|---|---|---|
| | RNN | LSTM | GRU | Avg. | RNN | LSTM | GRU | Avg. |
| Noun | 57.14 | 76.19 | 75.24 | 69.52 | 56.36 | 74.55 | 74.07 | 68.33 |
| Noun Adjective | 60.38 | 80.19 | 74.53 | 71.70 | 58.93 | 76.27 | 82.05 | 72.42 |
| Noun Adjective verb | 68.87 | 77.36 | 77.36 | 74.53 | 67.27 | 71.88 | 78.00 | 72.38 |
| All parts of speech | 62.26 | 77.36 | 73.58 | 71.07 | 58.82 | 71.88 | 65.38 | 65.36 |
| Average | 62.16 | 77.78 | 75.18 | | 60.35 | 73.65 | 74.88 | |

Step 1. **morpheme analysis**

Konlpy API - okt

Step 2**. comparing 3 NN model**

RNN, LSTM, GRU

- Get n. / n. + adj / n.+ adj.+v. / all
- **13K** comments from 10 news on same topic.
- Toxic comments rate = 2 %
- LSTM is Greatest in accuracy ( **80.19%**)

Jin Woo Kim, A Comparison Study on Performance of Malicious Comment Classification Models Applied with Artificial Neural Network, 2019

# My Research

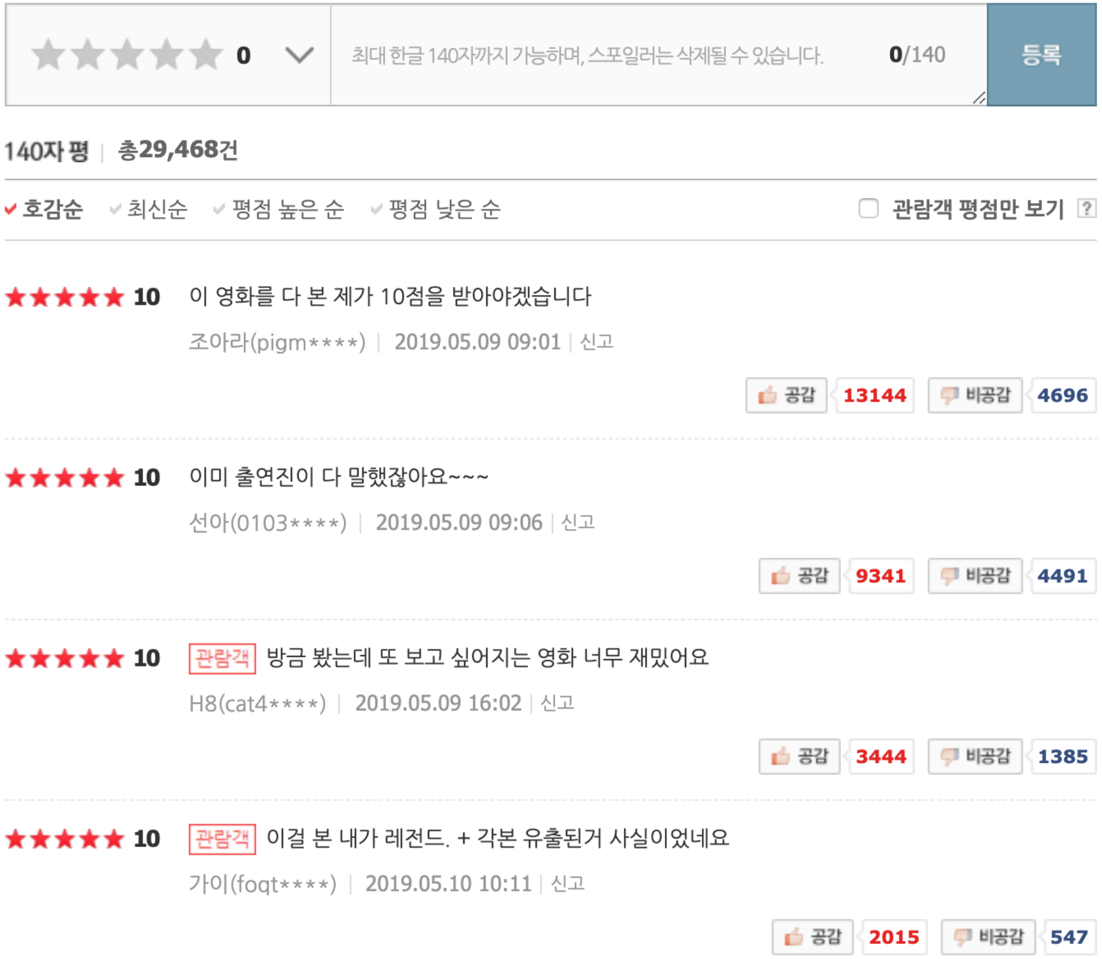
# Research Flow

1. **Collect the data** - from Naver Movie
2. **Labeling the data** - CrowdSourcing
3. **Make Korean word vector** – by character level & jamo level
4. **Implement text classification model** – textCNN
5. **Evaluation**

Collect the data

# Crawling

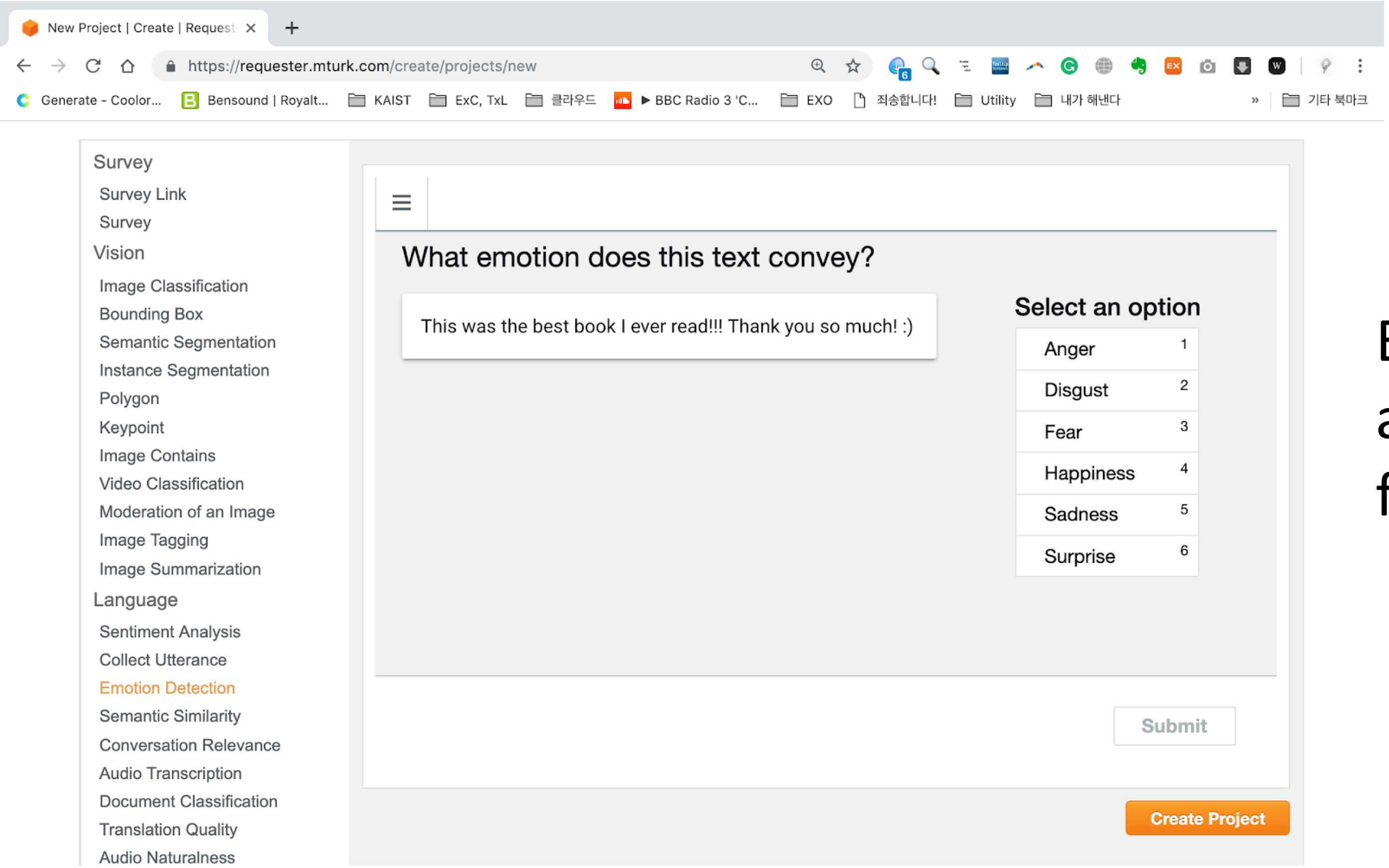
From 15 movies, collected 120K comments




| Movie Title | Genre | # of comments |
|---|---|---|
| 걸캅스 | 코미디/액션 | 27972 |
| 보헤미안 | 드라마 | 38846 |
| 덩케르크 | 액션/드라마/스릴러/전쟁 | 17289 |
| 캡틴마블 | 액션/모험/SF | 35260 |
| 주토피아 | 애니메이션/액션/모험/코미디/가족 | 17680 |
| 님아, 그 강을 건너지 마오 | 다큐멘터리 | 14651 |
| 아가씨 | 스릴러/드라마 | 22785 |
| 검은사제들 | 미스터리/드라마 | 20452 |
| 위대한 쇼맨 | 드라마/뮤지컬 | 13380 |
| 겨울왕국 | 애니메이션/모험/코미디/가족/판타지/뮤지컬 | 35125 |
| 다크나이트 | 액션/범죄/드라마/미스터리 | 26156 |
| 아바타 | SF/모험/액션/전쟁 | 40992 |
| 아이언맨 | SF/액션/드라마/판타지 | 10566 |
| 하울의 움직이는성 | 애니메이션/판타지 | 11552 |
| 타이타닉 | 멜로/로맨스/드라마 | 20260 |

# Label the data

## Crowdsourcing

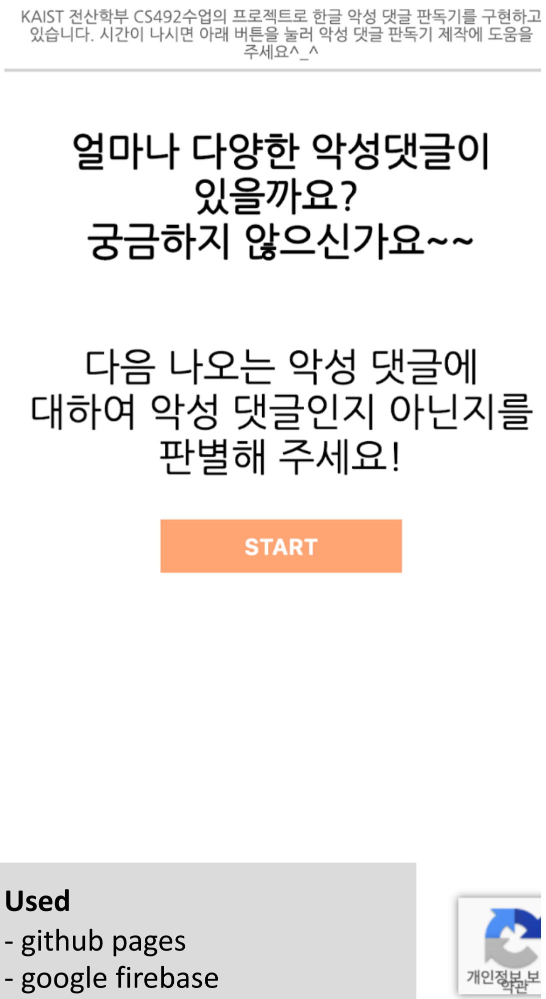



Existing Crowdsourcing platforms are **NOT easily accessible** for **Korean Users**

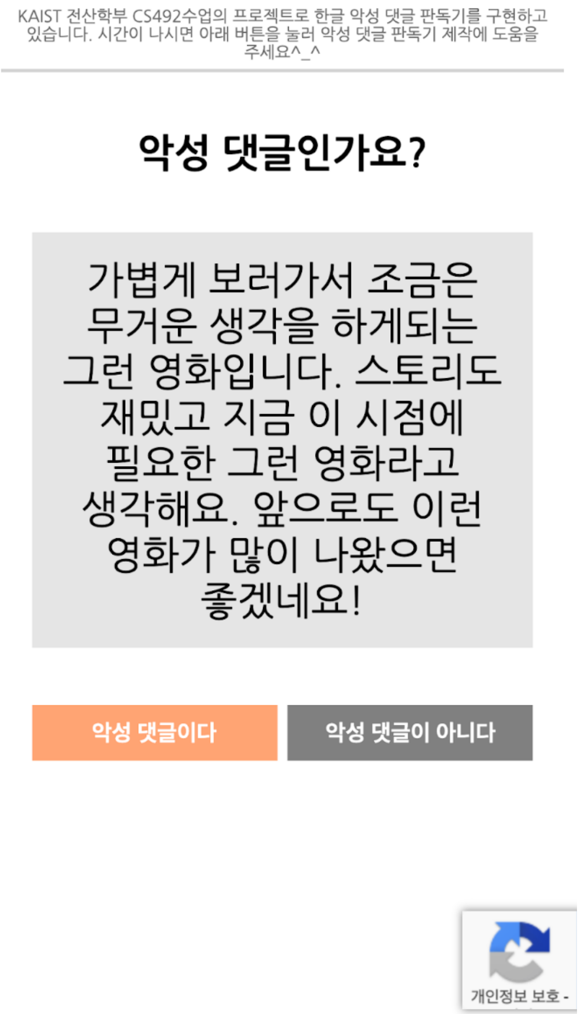
MTURK: Amazon data crowdsourcing site

# Label the data

# Crowdsourcing









**Used**
- github pages
- google firebase

crowdsourcing

# Collected Data

| Movie Title | Total | Toxic |
|---|---|---|
| My Love | 236 | 39 |
| Priests | 246 | 21 |
| Showman | 257 | 14 |
| Titanic | 284 | 40 |
| Zootopia | 146 | 12 |
| Avatar | 267 | 33 |
| Bohemian | 263 | 16 |
| Captain | 296 | 68 |
| Darkknight | 300 | 53 |
| Dunkirk | 187 | 22 |
| Frozen | 257 | 14 |
| Girlcops | 312 | 63 |
| Handmaiden | 181 | 79 |
| Howl | 209 | 14 |
| Ironman | 241 | 13 |
| 15 movies | 3682 | 501 |

**Total 3682 comments**
**13% toxic comments**

**Sample toxic comments**

- 강동원 빠순이들이 전우치랑 비교하네 **ㅉㅉ**
- 시간이 아까운 게이영화
- ?? 이거 ㅋㅋ 댓글에 10점따리 준놈들 내용 읽어보니깐 **쿵쾅쿵쾅** 분들인거 다 알겠는데? ㅋㅋ 왠일로 10점이 많지? 했는데 역시.. 쿵쾅쿵쾅분들땜에 보기싫어서 안볼렵니다
- 1점준 놈들은 자전차왕 **읍읍읍** 개꿀잼일테니 보러가세요
- 도태돼서 안 보고 방구석에서 발광하는 거잖스 ㅎㅋ
- 박평식 저 관심병환자새끼는 그냥 묶어놓고 패야된다
- 10점 주신분들 영혼만 보내서 감명깊게보고 n차관람 하신다는거죠?
- 10점은 오바고 8점짜리 준수한 영화. 이게 라푼젤보다 잘만들었다고 하는건 이해불가. 스토리 자체가 라푼젤에 비해선 한참 딸림.
- 개꿀잼이다 형은 **이번에세번째로보러간다**. 엘사갓찬양해 별점낮은애들은걍.집에서 토렌트로본그지념들임 ㅉㅉ 개명작
- 황정민 오달수 이경영같이 믹스커피+담배냄새 절거같은 아재들만 주구장창 보다가 이런 꽃미남들이 단체로 나오는 영화봐서 좋았습니다! 눈정화 잇힝^^
- 개봉되서는 언될 전형 작품성도 없고 최악의 쓰레기 영화네요
- 초딩들을 위한영화 3류영화다

# Preprocess the data

1. Make continuous punctuation marks(?!,.) into one **punctuation mark**.

2. **Do proper spacing** which have more than 5 characters.

3. **Remove useless comments** such as containing only hyperlinks or any random characters
   -> 5 comments were removed.

Fahim Mohammad, “Is preprocessing of text really worth your time for toxic comment classification?, 2018

Make word vector

# Is preprocessing of text really worth your time for toxic comment classification?(2018)

| | F1-score | | | | Overall Accuracy | | | | Total Misclassified Comments (out of 159580) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Preprocessing Steop** | **Logit** | **NBSVM** | **fastText-BiL** | **XGBoost** | **Logit** | **NBSVM** | **fastText-BiL** | **XGBoost** | **Logit** | **NBSVM** | **fastText-BiL** | **XGBoost** |
| **Raw** | 0.7407 | 0.7957 | 0.7906 | 0.5727 | 0.9720 | 0.9773 | 0.9775 | 0.9196 | 6992 | 5946 | 6217 | 9864 |
| **To_lower** | 0.7352 | 0.7936 | 0.8055 | 0.5757 | 0.9715 | 0.9773 | 0.9787 | 0.9196 | 7112 | 5974 | 5882 | 9809 |
| **Remove_whitespaces** | 0.7407 | 0.7957 | 0.7903 | 0.5727 | 0.9720 | 0.9773 | 0.9774 | 0.9196 | 6992 | 5946 | 6206 | 9864 |
| **Remove_leaky** | 0.7405 | 0.7951 | 0.7916 | 0.5735 | 0.9721 | 0.9773 | 0.9766 | 0.9198 | 7000 | 5958 | 6205 | 9849 |
| **trim_words_len** | 0.7401 | 0.7958 | 0.7906 | 0.5726 | 0.9720 | 0.9773 | 0.9774 | 0.9197 | 7009 | 5946 | 6230 | 9860 |
| **Strip_non_printables** | 0.7402 | 0.7959 | 0.7947 | 0.5729 | 0.9719 | 0.9772 | 0.9778 | 0.9197 | 7005 | 5940 | 6168 | 9863 |
| **Replace_contractions** | 0.7399 | 0.7951 | 0.7885 | 0.5736 | 0.9720 | 0.9773 | 0.9769 | 0.9201 | 7014 | 5965 | 6242 | 9844 |
| **Replace_acronyms** | 0.7393 | 0.7934 | 0.7876 | 0.5738 | 0.9719 | 0.9769 | 0.9775 | 0.9198 | 7038 | 6003 | 6287 | 9879 |
| **Remove_stopwords** | 0.7302 | 0.7860 | 0.7904 | 0.5643 | 0.9706 | 0.9733 | 0.9773 | 0.9007 | 7186 | 6209 | 6237 | 10013 |
| **Remove_rare_words** | 0.7297 | 0.7849 | 0.7735 | 0.5608 | 0.9681 | 0.9719 | 0.9737 | 0.9142 | 7257 | 6243 | 6556 | 10048 |
| **Remove_non_alnum_chars** | 0.7307 | 0.7885 | 0.8028 | 0.5680 | 0.9705 | 0.9761 | 0.9791 | 0.9163 | 7199 | 6105 | 5935 | 9935 |
| **Remove_non_alpha_chars** | 0.7337 | 0.7905 | 0.8040 | 0.5697 | 0.9709 | 0.9762 | 0.9796 | 0.9165 | 7145 | 6068 | 5897 | 9905 |
| **Remove_non_alpha_words** | 0.6577 | 0.7084 | 0.7208 | 0.4824 | 0.9462 | 0.9481 | 0.9549 | 0.8866 | 8744 | 8012 | 7859 | 11196 |
| **Regex_maping_black_list** | 0.7488 | 0.7913 | 0.8006 | 0.6252 | 0.9736 | **0.9775** | 0.9796 | 0.9303 | 6854 | 6081 | 5950 | 9083 |
| **Check_if_name** | 0.7407 | 0.7957 | 0.7947 | 0.5727 | 0.9720 | 0.9773 | 0.9774 | 0.9196 | 6992 | 5946 | 6121 | 9864 |
| **Fuzzy_profane_map** | 0.7422 | 0.7855 | 0.7910 | 0.6082 | 0.9718 | 0.9753 | 0.9775 | 0.9258 | 6999 | 6223 | 6293 | 9342 |
| **Fuzzy_common_map** | 0.7438 | **0.7968** | 0.7914 | 0.5794 | 0.9724 | 0.9769 | 0.9774 | 0.9224 | 6933 | **5933** | 6227 | 9758 |
| **Lemmatize** | 0.7377 | 0.7888 | 0.7918 | 0.5722 | 0.9698 | 0.9734 | 0.9774 | 0.9194 | 7091 | 6126 | 6208 | 9877 |
| **Stemming** | 0.7322 | 0.7782 | 0.8023 | 0.5919 | 0.9683 | 0.9715 | 0.9794 | 0.9225 | 7216 | 6390 | 5878 | 9568 |
| **URL_info_extract** | 0.7396 | 0.7953 | 0.7828 | 0.5735 | 0.9719 | 0.9773 | 0.9776 | 0.9199 | 7016 | 5958 | 6274 | 9853 |
| **PPO-1-lower_ws_trim** | 0.7351 | 0.7934 | 0.8006 | 0.5735 | 0.9715 | 0.9773 | 0.9783 | 0.9195 | 7113 | 5979 | 5955 | 9845 |
| **PPO-2-LWTN-Lk** | 0.7366 | 0.7926 | 0.7994 | 0.5738 | 0.9716 | 0.9773 | 0.9789 | 0.9193 | 7078 | 6003 | 5947 | 9839 |
| **PPO-3-LWTN-LkCnAc** | 0.7311 | 0.7825 | 0.7961 | 0.5689 | 0.9709 | 0.9760 | 0.9777 | 0.9195 | 7232 | 6281 | 6071 | 9986 |
| **PPO-4-LWTN-St** | 0.7247 | 0.7826 | 0.7970 | 0.5641 | 0.9702 | 0.9734 | 0.9783 | 0.9007 | 7291 | 6266 | 5994 | 10010 |
| **PPO-5-LWTN-Ra** | 0.7298 | 0.7846 | 0.7932 | 0.5641 | 0.9690 | 0.9733 | 0.9756 | 0.9159 | 7237 | 6216 | 6172 | 9996 |
| **PPO-6-LWTN-CoAcStRa** | 0.7148 | 0.7647 | 0.7830 | 0.5538 | 0.9660 | 0.9672 | 0.9733 | 0.8958 | 7569 | 6754 | 6433 | 10251 |
| **PPO-7-LWTN-An** | 0.7240 | 0.7844 | 0.8076 | 0.5694 | 0.9699 | 0.9761 | 0.9801 | 0.9157 | 7331 | 6170 | 5756 | 9908 |
| **PPO-8-LWTN-Aw** | 0.7278 | 0.7859 | **0.8117** | 0.5724 | 0.9702 | 0.9762 | 0.9795 | 0.9161 | 7260 | 6139 | **5715** | 9862 |
| **PPO-9-LWTN-AnAw** | 0.7278 | 0.7859 | 0.8079 | 0.5724 | 0.9702 | 0.9762 | **0.9802** | 0.9161 | 7260 | 6139 | 5751 | 9862 |
| **PPO-10-LWTN-CoAcBk** | 0.7421 | 0.7815 | 0.7995 | 0.6236 | 0.9727 | 0.9763 | 0.9780 | 0.9306 | 7024 | 6327 | 6043 | 9161 |
| **PPO-11-LWTN-CoAcBkPrCm** | 0.7466 | 0.7790 | 0.7993 | 0.6302 | 0.9733 | 0.9759 | 0.9778 | 0.9325 | 6944 | 6404 | 6103 | 9075 |
| **PPO-12-LWTN-CoAcLkBkPrCmNm** | 0.7477 | 0.7792 | 0.8004 | 0.6305 | 0.9733 | 0.9759 | 0.9775 | 0.9322 | 6922 | 6399 | 6089 | 9074 |
| **PPO-13-LWTN-CoAcLkAwStSm** | 0.7292 | 0.7680 | 0.8009 | 0.5871 | 0.9693 | 0.9709 | 0.9778 | 0.9123 | 7299 | 6649 | 6005 | 9752 |
| **PPO-14-lower_lemma** | 0.7338 | 0.7868 | 0.8038 | 0.5721 | 0.9701 | 0.9743 | 0.9787 | 0.9208 | 7163 | 6150 | 5900 | 9876 |
| **PPO-15-lower-AwBkCmSm** | **0.7519** | 0.7884 | 0.8076 | **0.6348** | **0.9739** | 0.9768 | 0.9786 | **0.9327** | **6816** | 6139 | 5877 | **8919** |

Fig. 4: Results: F1 scores, accuracies and total number of misclassified.

Fahim Mohammad, "Is preprocessing of text really worth your time for toxic comment classification?, 2018

Make word vector

# Make Korean word vector using fasttext

## 1. By character level

- use basic fasttext skipgram.
- build word vector by considering meaning per characters(유,치,뽕,영, 화)

## 2. By jamo level

- Decompose word by jamo and learn by fasttext skipgram.

ㅇㅠeㅊㅣeㅃㅗㅇ ㅇㅕㅇㅎㅗㅏe

- better for semantic and syntactic similarity and analogy tasks.

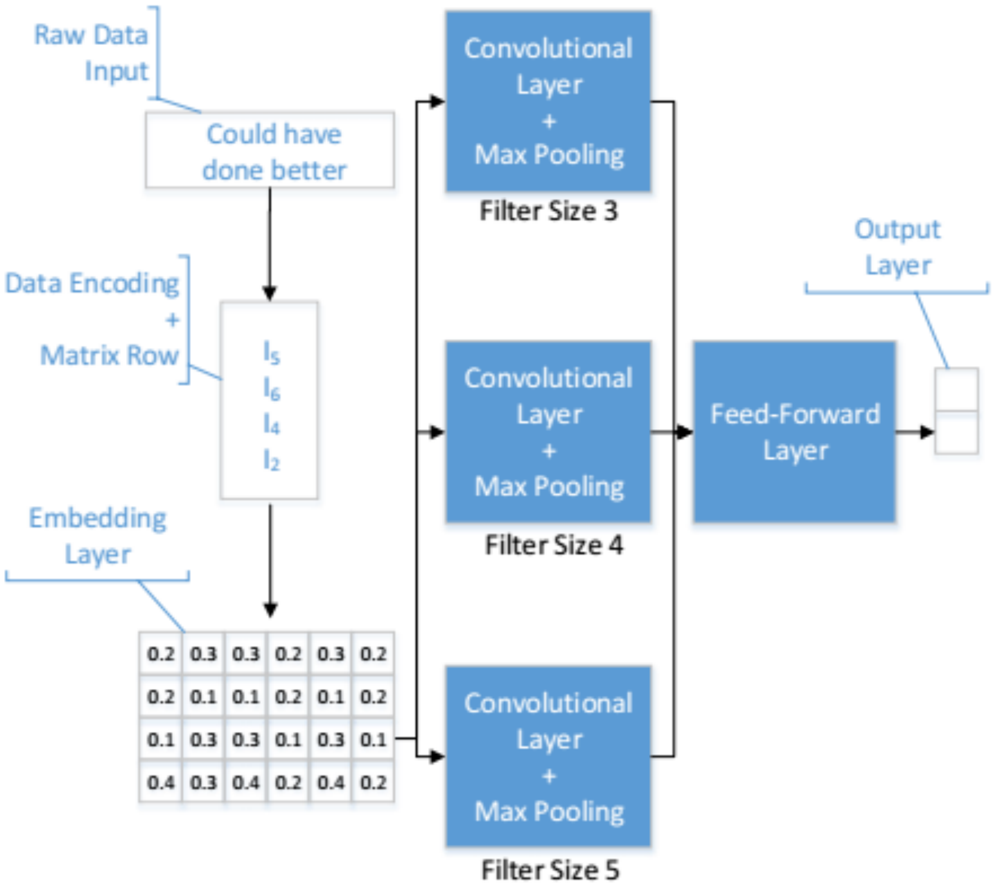
based on Park et al., 2018b S. Park, Byun J., S. Baek, Cho Y., A. Oh "Subword-level word vector representations for Korean Proceedings of the fifty-sixth annual meeting of the association for computational linguistics (2018)", pp. 2429-2438

CNN model

# textCNN model(2018)

- Recently, CNN are being applied to text classification or NLP **without using syntactic or semantic knowledge of a language**.

- CNN using character-level feature is effective method.




based on Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis and Vassilis P. Plagianakos, 2018, "Convolutional Neural Networks for Toxic Comment Classification" Proceedings of the 10th Hellenic Conference on Artificial Intelligence Article No. 35

CNN model

# textCNN for toxic classification(2018)

**Table 1: Mean values and Standard Deviation across all experiments for Accuracy, Specificity and False discovery rate for all Classification Methods.**

| | Accuracy | | Specificity | | False disc.rate | |
|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std |
| $CNN_{fix}$ | 0.912 | 0.002 | 0.917 | 0.006 | 0.083 | 0.007 |
| $CNN_{rand}$ | 0.895 | 0.003 | 0.906 | 0.015 | 0.092 | 0.017 |
| kNN | 0.697 | 0.008 | 0.590 | 0.016 | 0.335 | 0.010 |
| LDA | 0.808 | 0.005 | 0.826 | 0.010 | 0.179 | 0.009 |
| NB | 0.719 | 0.005 | 0.776 | 0.012 | 0.250 | 0.010 |
| SVM | 0.811 | 0.007 | 0.841 | 0.012 | 0.167 | 0.012 |

based on Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis and Vassilis P. Plagianakos, 2018, "Convolutional Neural Networks for Toxic Comment Classification” Proceedings of the 10th Hellenic Conference on Artificial Intelligence Article No. 35

# Evaluation

Evaluation

# Results

| | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| Jamo + CNN | 0.695 | 0.704 | 0.708 | 0.700 |
| Character + CNN | **0.813** | **0.754** | **0.858** | **0.801** |

- Average of **5** times random train-test split
- **Test set** civil: toxic = 1: 1
- **Train set** civil: toxic = 1.1: 1

Parameters used

- Text- CNN: Embedding_dim =300, dropout_keep_rate = 0.85, dev_sample = 10%,
  l2_regularization = 1.0, batch_size = 100, num_epochs= 2500
- Fasttext by jamo-level: skipgram, minCount=1, minjn=3, maxjn=5, minn=1, maxn=4, dim=300, ws=5
- Fasttext by character-level: skipgram, minCount=1, minn=1, maxn=4, dim=300, ws=5

# Evaluation

| | **Accuracy** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|---|
| Jamo + CNN | 0.695 | 0.704 | 0.708 | 0.700 |
| Character + CNN | **0.813** | **0.754** | **0.858** | **0.801** |

1. First attempt to use Korean word-vector and CNN model.
2. Great performance although it has small data(3k < 13k< 160k).

# Future Work

## For Better Accuracy

1. Get more labeled data
2. Adjust the threshold of toxicity from crowdsourcing.
3. Adopt Pseudo Labeling Method
4. More work on constructing word-vector & CNN parameter fitting

# Future Work

**After getting enough accuracy and data,**

1. Follow up advanced English Hate speech research.
2. Research on data from Korean SNS platform

# Application

# Application

1. **Alert** that specific chunks can make people feel bad.

# Application

2. **Predict future** conversation and avoid the fight.

3. Build a **typology of antisocial users**

conclusion

# remaining problems..

Thanks for listening ~ ^_^!