Jibes & Delights: A Dataset of Targeted Insults and Compliments to Tackle Online Abuse

Ravsimar Sodhi and **Kartikey Pant** and **Radhika Mamidi**
International Institute of Information Technology
Hyderabad, Telangana, India
ravsimar.sodhi@research.iiit.ac.in
kartikey.pant@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract

Online abuse and offensive language on social media have become widespread problems in today's digital age. In this paper, we contribute a Reddit-based dataset, consisting of $68,159$ insults and $51,102$ compliments targeted at individuals instead of targeting a particular community or race. Secondly, we benchmark multiple existing state-of-the-art models for both classification and unsupervised style transfer on the dataset. Finally, we analyse the experimental results and conclude that the transfer task is challenging, requiring the models to understand the high degree of creativity exhibited in the data.

1 Introduction

Online abuse and targeted negative comments on online social networks have become a prevalent phenomenon, especially impacting adolescents. Victims of prolonged and targeted online harassment can experience negative emotions leading to adverse consequences such as anxiety and isolation from the community, which can, in turn, lead to suicidal behaviour.[1] Various attempts have been made to detect such harassment (Dadvar et al., 2013; Chatzakou et al., 2019) and hate speech (Davidson et al., 2017; Badjatiya et al., 2017) in the past, but very few have attempted to transfer the negative aspect of such speech.

Recently, many new tasks have been introduced in the domain of text style transfer. However, since parallel corpora is usually not available, most style transfer approaches adopt an unsupervised manner (Li et al., 2018; Zhang et al., 2018; John et al., 2019; Wang et al., 2019) . We contribute a dataset of non-parallel sentences, each sentence being either an insult or a compliment, collected from Reddit, more specifically, from three subreddits r/RoastMe, r/ToastMe, and r/FreeCompliments. Some examples of such sentences can be seen in Figure 1.

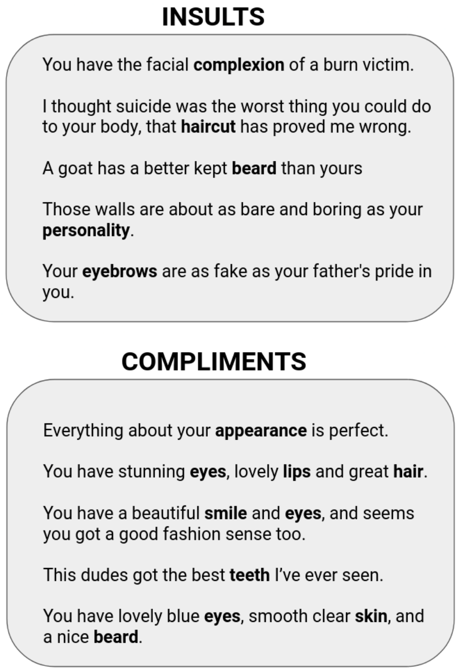

Figure 1: Examples of indirect insults and compliments with attributes highlighted in bold

With a diverse range of online communication platforms being introduced across the world, and existing platforms' user-bases growing at a fast pace, moderation of such negative comments and harassment becomes even more necessary. We hope that our work can enable and further research in this daunting task, for instance in building moderation systems which can detect such negative speech and nudge users to engage in more positive and non-toxic discourse.

Reddit is a popular social media website with forums known as subreddits where users can comment and vote on posts and other comments. It has been used as a source of data in wide variety of tasks. r/RoastMe can be described as a sub-

[1] Source: https://www.stopbullying.gov/resources/facts

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 132–139
August 6, 2021. ©2021 Association for Computational Linguistics

reddit consisting of "abrasive humor" and consists of "creative" insults where users can voluntarily submit their picture to be "roasted." r/ToastMe and r/FreeCompliments are similar in principle but have the opposite purpose. A more detailed description of the data source and preparation can be found in Section 3. Since "creativity" is encouraged in r/RoastMe, this makes our dataset consist of indirect insults that do not necessarily use any profanity or curse words and may slip past most existing toxic speech filters.

In this work, we release the JDC (**J**ibe and **D**elight **C**orpus), a dataset of $\sim 120,000$ Reddit comments tagged as insults or compliments which are targeted towards particular attributes of an individual including their *face*, *hair*, and *eyes*[2]. We also propose to use classification models to detect and style transfer models to convert such targeted negative comments, often associated with online harassment, in which menacing or insulting messages are sent by direct messages or posted on social media. We also perform benchmarking experiments using existing state-of-the-art models on both fronts and analyse its results.

2 Related Work

Existing work primarily focuses on the *detection* of offensive language or hate speech on social media using classification models (Davidson et al., 2017; Badjatiya et al., 2017; Dadu and Pant, 2020), and not on *transferring* the negative aspect of such speech into a positive counterpart. Detection usually involves either lexical or rule-based approach (Pérez et al., 2012; Serra and Venter, 2011), or more recently, a supervised learning approach (Yin et al., 2009; Dinakar et al., 2011). Many attempts on detection of specific types of toxic speech have also been attempted (Basile et al., 2019; Zampieri et al., 2020). Previous work on text style transfer has largely focused on transferring attributes of sentiment in reviews (Li et al., 2018; Hu et al., 2017; Pant et al., 2020) or converting factual captions to humorous or romantic ones (Li et al., 2018). Other tasks include transferring formality (Xu et al., 2019) or gender or political style (Reddy and Knight, 2016). Recently, transferring politeness has also been proposed by Madaan et al. (2020).

Most approaches use unsupervised methods since parallel data is usually not available. These approaches can be broadly divided into three groups: 1) *Explicit disentanglement* (Li et al., 2018; Sudhakar et al., 2019) which separates content from style attributes in an explicit manner and then combines the separated content with the target attribute and pass it through a generator. 2) *Disentanglement in latent space* (John et al., 2019) which tries to separate style from content within the embedding space by using suitable objective functions. 3) *Adversarial or reinforcement learning based* (Luo et al., 2019) approaches in which disentanglement may not be even required.

Reddit has been widely used in multiple natural language processing tasks as a data source. While Khodak et al. (2018) use Reddit to create a large corpus for sarcasm, Nogueira dos Santos et al. (2018) source their data from r/Politics on Reddit along with Twitter. Many controversial subreddits such r/The_Donald have been used for detection of hate speech in the past (Qian et al., 2019).

Although Nogueira dos Santos et al. (2018) proposed the task of translating offensive sentences to non-offensive ones using style transfer, in our work, we go one step further and propose to convert offensive sentences into positive compliments. Prior work on r/RoastMe has mostly been on a socio-pragmatic perspective (Dynel and Poppi, 2019; Kasunic and Kaufman, 2018). However, there is no previous work that uses r/RoastMe as a data source in a style transfer task to the best of our knowledge.

3 The JDC Dataset

We contribute the Jibe and Delight Corpus (JDC), a new non-parallel style transfer dataset consisting of $\sim 120,000$ comments tagged as insults or compliments, and perform experiments and analysis on the same.

3.1 Data Collection

We use Pushshift (Baumgartner et al., 2020) to extract Reddit posts and comments.While r/RoastMe is often characterized as a humourous subreddit, where users can voluntarily submit pictures of themselves to be "roasted" or insulted, internet users who are not familiar with the community can associate r/RoastMe with malicious activities including cyberbullying (Jenaro et al., 2018). r/RoastMe has even been described as "a new cy-

[2] Made available at https://github.com/ravsimar-sodhi/jibes-and-delights

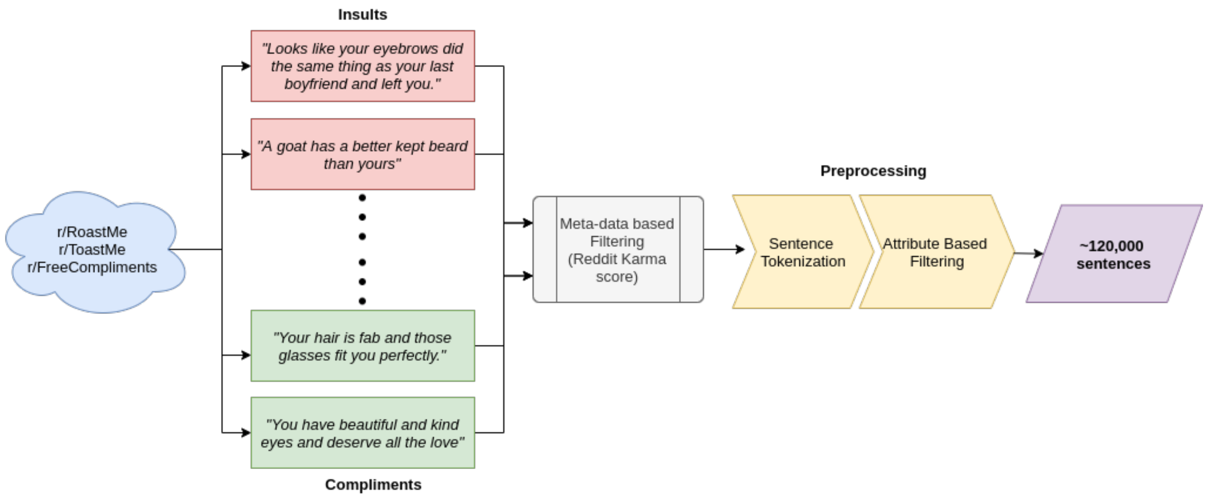

Figure 2: Dataset Creation Pipeline

berbullying trend" by news media[3]. The "roasters" will come up with comments that insult or demean the poster of the picture while trying to be as "creative" as possible. r/ToastMe and r/FreeCompliments work similarly, but with the opposite intent. These communities are much smaller and less popular than r/RoastMe, and hence, the number of insults in our dataset is greater than the number of compliments.

It is essential to distinguish between insults and slurs since the two are frequently clubbed together. A slur is a taboo remark that is usually used to deprecate, disparage or derogate a targeted member of a select category, such as ethnicity, race, or sexual orientation. Insults can consist of slurs, but they are much broader, and a more diversified phenomenon (Dynel and Poppi, 2019). While slurs are common in hate speech datasets, it is interesting to observe that slurs are comparatively rare in posts from r/RoastMe. This characteristic makes our dataset unique concerning other offensive speech datasets. Figure 2 illustrates our pipeline for the creation of the JDC.

3.2 Data Preparation

Since Reddit is a conversational social media platform, we limit the JDC to comments having the following characteristics:

- *They are top-level comments*. While nested comments may also sometimes contain relevant data, they often diverge from the topic and begin a conversation with the parent comment, which may add noise to our dataset.
- *They have a Reddit karma score of at least 3*. Reddit karma score is defined to be the number of upvotes minus the number of downvotes. This filtering helps in weeding out spam or irrelevant comments which are not relevant to the topic. Generally, users downvote comments which they find unfunny or off-topic. Thus, we utilize crowdsourced user scores to ensure quality sentences.

This filtering yields a corpus of roughly $300,000$ insult comments and $100,000$ compliment comments. However, due to users' wide variety of insults and creativity, we utilize several other filters to limit our dataset to a particular type of insult or compliment. We manually create an attribute list[4] consisting of words which correspond to a physical attribute (for example, *hair*, *skin*, *complexion*, *teeth*, and *eyes*) or a trait (for example, *personality*, *kindness*, and *appearance*) and keep our insults containing keywords for such attributes. We use the NLTK (Bird et al., 2009) and spaCy (Honnibal and Montani, 2017) libraries for preprocessing and filtering. We tokenize each comment into sentences, and check if the lemma of any word in a sentence matches a word in our attribute list. This process helps us keep relevant sentences, especially from longer comments, which would be otherwise discarded. We also filter out very short sentences (containing only one or two words). We obtain

[3]Source:https://abcnews.go.com/Lifestyle/parents-roasting-cyberbullying-trend/story?id=49407671

[4]https://github.com/ravsimar-sodhi/jibes-and-delight/blob/main/attr-list.txt

Input	Your teeth are more stained than my toilet
StyleEmb	Your hair is beautiful than your face.
RetrieveOnly	Beautiful smile - your teeth are remarkably straight
DeleteRetrieve	Your hair is beautiful and your teeth are more stained than my
LingST	Your teeth are so beautiful
Tag&Gen	Your teeth are more stained than my heart
Input	The only thing lazier than your eye is God when he designed your busted face
StyleEmb	Keep your hair and I love your hair and you look like the kind of person who look like a
RetrieveOnly	Your hair is fantastic and your face is absolutely adorable.
DeleteRetrieve	And your eye expressive is God wonderful lips designed especially your face
LingST	The only thing more crooked than your face is the absolute cutest thing i
Tag&Gen	Love the only thing lazier than your eye is god when he designed your busted face

Table 1: Examples of Style Transfer model outputs

a corpus of around $68,159$ *insult* sentences and $51,102$ *compliment* sentences. Finally, we take $1,000$ instances each from both categories for evaluation purposes.

4 Experiments & Results

We perform experiments using both classification and style transfer models and evaluate the performance of five models for each task from works on the JDC. We also discuss about the challenges faced and the metrics used for evaluation.

4.1 Models

For classification experiments, we experiment using the following models:

1. **Logistic Regression**: One of the most common classification algorithms used, Logistic Regression (*LR*) uses the logistic (sigmoid) function to return a probability value that can be further mapped to multiple classes.

2. **SVM**: Support Vector Machines (*SVM*) use an objective function that finds a hyperplane in an N dimensional space, where N is the number of features, which distinctly separates the data points into classes.

3. **BERT** (Devlin et al., 2019): Bidirectional Encoder Representations from Transformers or *BERT* is a relatively recent transformer-based model, which leverages transfer learning. At the time of release, *BERT* outperformed several other models in language modeling tasks.

4. **RoBERTa** (Liu et al., 2019): *RoBERTa* improves on *BERT* by modifying several hyperparameters and performs pretraining on larger amounts of data for a longer amount of time.

5. **XLNet** (Yang et al., 2019): While *BERT* and *RoBERTa* are categorized as autoencoder language models, *XLNet* is a generalized autoregressive pretraining method. Instead of using Masked Language modeling like *BERT*, it proposed a new objective called Permuation Language Modeling, and its results improved upon *BERT* in many tasks.

We fine-tune the models for classification in case of **BERT**, **RoBERTa**, and **XLNet**, and use the Hugging Face's transformers library (Wolf et al., 2020) for our experiments.

For style transfer experiments, we use the following models:

1. **StyleEmb** (Shen et al., 2017): This model uses a cross-aligned auto-encoder, aligning representations in latent space to perform style transfer.

2. **RetrieveOnly** and **DeleteRetrieve** (Li et al., 2018): While *RetrieveOnly* only returns a sentence from the target style without any changes, *DeleteRetrieve* returns the *best match* according to the attribute markers from the source sentences. Both models use explicit disentanglement to separate content from style along with a decoder and are often used as baselines in multiple works in style transfer.

3. **LingST** (John et al., 2019): This model uses a variational auto-encoder and utilizes multiple adversarial objectives for both style and content preservation.

4. **Tag&Gen** (Madaan et al., 2020): This model uses an encoder-decoder approach where both encoder and decoder are transformer-based.

Model	Acc.	Prec.	Recall	F1
LR	0.875	0.980	0.897	0.883
SVM	0.801	0.979	0.851	0.818
BERT	0.977	0.967	**0.974**	0.970
RoBERTa	0.977	**0.971**	0.973	0.971
XLNet	**0.978**	0.970	0.973	**0.972**

Table 2: Automatic Evaluation Results of Classification models on the dataset

This has recently been utilized for the politeness transfer task.

4.2 Evaluation

For the classification experiments, we evaluate using the well-known metrics of Accuracy, Precision, Recall and F1-Score.

For the style transfer experiments, we evaluate the performance on three different aspects, following previous works:

1. **Style Transfer Intensity**: We train a separate fastText model (Joulin et al., 2017) on the training data and evaluate the different model outputs to determine the accuracy of style transfer.

2. **Content Preservation**: We use BLEU as an evaluation metric and utilize the *SacreBLEU* (Post, 2018) implementation.

3. **Fluency**: We calculate fluency using a language model from the *KenLM* library for our experiments. (Heafield, 2011) after training the language model on the target domain (*compliment*). A lower perplexity indicates a more fluent sentence and vice-versa.

Apart from automatic evaluation, we also do human evaluation on 280 sentences randomly selected from the test set. The evaluators were asked to rank sentences on basis of their fluency and degree of being a compliment (DOC) on a scale of 1 to 5. Two annotators were shown a list of sentences, with no indication of the source of the sentence. The Cohen's Kappa metric (Cohen, 1960) was used to measure the agreement between the two annotators. The value of kappa for DOC and fluency come out to be 0.69 and 0.65 respectively.

4.3 Results

Table 2 shows that most of the models perform very well in classifying insults and compliments into different categories. Even ML-based models like *LR* and *SVM* perform adequately on the task, but the more state-of-the-art BERT-based models perform excellently, having high F1-scores above 0.9. *XLNet* shows the highest F1-score, with *BERT* and *RoBERTa* only marginally lower.

Model	Acc(%)	BLEU	PPL
StyleEmb	87.41	2.27	615.59
RetrieveOnly	**97.77**	3.83	241.04
DeleteRetrieve	81.35	23.81	857.19
LingST	93.00	3.16	**63.03**
Tag&Gen	30.17	**85.40**	637.39

Table 3: Automatic Evaluation results of Style Transfer models on the dataset

Model	DOC	Fluency
Input	1.116	**4.648**
StyleEmb	1.904	1.786
RetrieveOnly	**4.170**	4.468
DeleteRetrieve	2.595	2.051
LingST	3.851	3.414
Tag&Gen	1.382	3.819

Table 4: Human Evaluation results of Style Transfer models on the dataset. DOC is the "Degree of Compliment" and Fluency is the naturalness of the sentence, both being rated on a scale of 1 to 5

From Table 3, we observe that *RetrieveOnly*, *StyleEmb*, and *LingST* show high accuracy in transfer but do not perform well in content preservation. *Tag&Gen* performs very well on content preservation but fails to transfer the style adequately. *DeleteRetrieve* obtains a better balance in accuracy and BLEU, but it loses out on fluency, producing the least fluent sentences among all models. This implies that although the relevant words from style and content are transferred, the output may not be grammatical or natural.

Human evaluation results in Table 4 also show that the input sentence is judged as more fluent rather than the model outputs. We see that *RetrieveOnly* and *LingST* outputs are more likely to be judged as compliments. However, *StyleEmb* is judged to be as poor in both DOC and Fluency.

4.4 Discussion

Even though the insults are "creative", we find that the classification models perform excellently in differentiating insults and compliments into two separate categories. This shows that both insults

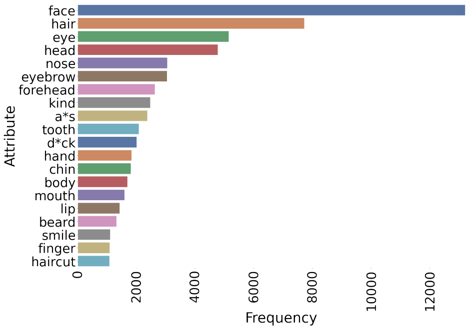

Figure 3: Insult distribution by top 20 attributes

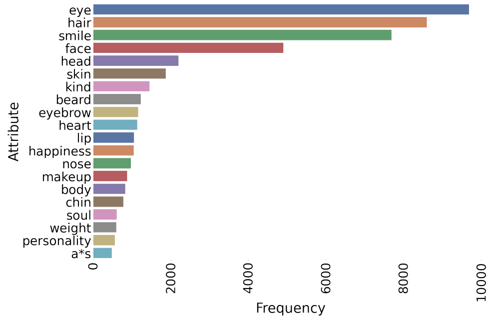

Figure 4: Compliment distribution by top 20 attributes

and compliments, even though they may target the same physical attributes, have different styles and models can be trained to detect and classify them with good results.

However, performance of the style transfer models is lacking. We observe that there are different shortcomings in each model. Table 1 shows sample outputs produced by the models. *RetrieveOnly* fetches a sentence from the target style using specific attributes from the input sentence, often leading to invalid outputs if a corresponding positive sentence does not exist in the data. In the second example in Table 1, *DeleteRetrieve* gives a nonsensical output. Other models have similarly tried to transfer the intent by introducing words like "kind","wonderful" and "cute", but there is still a significant gap between the generated outputs and genuine compliments. We observe that *LingST* has the lowest perplexity while still having high accuracy. This ensures that generated outputs are positive and are also grammatically fluent. Compared to sentiment transfer, converting an insult to a compliment usually involves multi-word modifications, explaining the poor content preservation across most of the models.

We observe that style transfer models have an easier time handling more direct insults ("You look very ugly"), rather than handling more complex and creative insults ("How many concrete walls did you have to run into to achieve that nose?"). Besides the more direct and creative the insults, there are some samples which need more context to understand and may seem out of place compared to the rest. However, most of these are filtered out with the help of the attribute list described in Section 3. The distribution of the data according to the top attributes can be seen in Figure 3 and Figure 4 for insults and compliments respectively. While a lot of creativity is exhibited in the insults, we find that compliments are usually of a direct form, and thus are simpler and easier to understand. This is desirable since some convoluted compliments may come across as patronizing, which is counterproductive to our goal.

5 Conclusion

In this work, we introduced a Reddit-based dataset consisting of indirect insults that favor creativity and rarely use slurs. We benchmarked classification models for detection and exploited unsupervised text style transfer to convert insults into compliments. We evaluated the performance of different state-of-the-art models on the dataset, observing that while detection is easier, transfer of the negative attribute is a challenging task. Future work may include enhancing methodologies for unsupervised text style transfer that capture the intricacies in the proposed dataset and building moderation systems for online platforms.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:830–839.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*, 1st edition. O'Reilly Media, Inc.

Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, Athena Vakali, and Nicolas Kourtellis. 2019. Detecting Cyberbullying and Cyberaggression in Social Media. *ACM Transactions on the Web*, 13(3):17:1–17:51.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Tanvi Dadu and Kartikey Pant. 2020. Team rouges at SemEval-2020 task 12: Cross-lingual inductive transfer to detect offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2183–2189, Barcelona (online). International Committee for Computational Linguistics.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving Cyberbullying Detection with User Context. In *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 693–696, Berlin, Heidelberg. Springer.

Thomas Davidson, Dana Warmsley, M. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

K. Dinakar, Roi Reichart, and H. Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. In *The Social Mobile Web*.

Marta Dynel and Fabio I. M. Poppi. 2019. Risum teneatis, amici?: The socio-pragmatics of RoastMe humour. *Journal of Pragmatics*, 139:1–21.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, WMT '11, pages 187–197, USA. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Zhiting Hu, Zichao Yang, Xiaodan Liang, R. Salakhutdinov, and E. Xing. 2017. Toward Controlled Generation of Text. In *ICML*.

Cristina Jenaro, Noelia Flores, and Cinthia Patricia Frías. 2018. Systematic review of empirical studies on cyberbullying in adults: What we know and what we should investigate. *Aggression and Violent Behavior*, 38:113–122.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Florence, Italy. Association for Computational Linguistics.

Armand Joulin, E. Grave, P. Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *EACL*.

Anna Kasunic and Geoff Kaufman. 2018. "At Least the Pizzas You Make Are Hot": Norms, Values, and Abrasive Humor on the Subreddit r/RoastMe. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

M. Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A Large Self-Annotated Corpus for Sarcasm. *LREC*.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. *NAACL-HLT*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*.

Fuli Luo, Peng Li, J. Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and X. Sun. 2019. A Dual Reinforcement Learning Framework for Unsupervised Text Style Transfer. *IJCAI*.

Aman Madaan, Amrith Rajagopal Setlur, Tanmay Parekh, B. Póczos, Graham Neubig, Yiming Yang, R. Salakhutdinov, A. Black, and Shrimai Prabhumoye. 2020. Politeness Transfer: A Tag and Generate Approach. In *ACL*.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia. Association for Computational Linguistics.

Kartikey Pant, Yash Verma, and Radhika Mamidi. 2020. SentiInc: Incorporating Sentiment Information into Sentiment Transfer Without Parallel Data. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Advances in Information Retrieval*, volume 12036, pages 312–319. Springer International Publishing, Cham.

Perla Janeth Castro Pérez, Christian Javier Lucero Valdez, María De Guadalupe Cota Ortiz, Juan Pablo Soto Barrera, and Pedro Flores Pérez. 2012. MISAAC: Instant messaging tool for ciberbullying detection. In *Proceedings of the 2012 International Conference on Artificial Intelligence, ICAI 2012*, Proceedings of the 2012 International Conference on Artificial Intelligence, ICAI 2012, pages 1049–1052.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding-Royer, and William Yang Wang. 2019. A Benchmark Dataset for Learning to Intervene in Online Hate Speech. In *EMNLP/IJCNLP*.

S. Reddy and K. Knight. 2016. Obfuscating Gender in Social Media Writing. In *NLP+CSS@EMNLP*.

S. M. Serra and H. S. Venter. 2011. Mobile cyberbullying: A proposal for a pre-emptive approach to risk mitigation by employing digital forensic readiness. In *2011 Information Security for South Africa*, pages 1–5.

T. Shen, Tao Lei, R. Barzilay, and T. Jaakkola. 2017. Style Transfer from Non-Parallel Text by Cross-Alignment. In *NIPS*.

A. Sudhakar, Bhargav Upadhyay, and A. Maheswaran. 2019. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. *EMNLP/IJCNLP*.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable Unsupervised Text Attribute Transfer via Editing Entangled Latent Representation. In *NeurIPS*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality Style Transfer with Hybrid Textual Annotations. *ArXiv*.

Z. Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. XL-Net: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian Davison, April Edwards, and Lynne Edwards. 2009. Detection of harassment on Web 2.0.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, M. Zhou, and E. Chen. 2018. Style Transfer as Unsupervised Machine Translation. *ArXiv*.