Error Analysis in a Hate Speech Detection Task: the Case of HaSpeeDe-TW at EVALITA 2018

Chiara Francesconi
Dipartimento di Lingue e Letterature Straniere e Culture Moderne
University of Turin
chiara.francesconi@edu.unito.it

Cristina Bosco
Fabio Poletto
Manuela Sanguinetti
Dipartimento di Informatica
University of Turin
{bosco,poletto,msanguin}@di.unito.it

Abstract

Taking as a case study the Hate Speech Detection task at EVALITA 2018, the paper discusses the distribution and typology of the errors made by the five best-scoring systems. The focus is on the sub-task where Twitter data was used both for training and testing (HaSpeeDe-TW). In order to highlight the complexity of hate speech and the reasons beyond the failures in its automatic detection, the annotation provided for the task is enriched with orthogonal categories annotated in the original reference corpus, such as aggressiveness, offensiveness, irony and the presence of stereotypes.

1 Introduction

The field of Natural Language Processing witnesses an ever-growing number of automated systems trained on annotated data and built to solve, with remarkable results, the most diverse tasks. As performances increase, resources, settings and features that contributed to the improvement are (understandably) emphasized, but sometimes little or no room is given to an analysis of the factors that caused the system to misclassify some items.

This paper wants to draw attention to the importance of a thorough error analysis on the performance of supervised systems, as a means to produce advancement in the field. Errors made by a system may entail not only the poorness of the system itself but also the sparseness of the data used in training, the failure of the annotation scheme in describing the observed phenomena or a cue of the data inherent ambiguity. The presence of the same errors in the results of several systems involved in a shared task may result in also more interesting hints about the directions to be followed in the improvement of both data and systems.

As a case study to carry out error analysis, data from a shared task have been used in this paper. Shared tasks offer clean, high-quality annotated datasets on which different systems are trained and tested. Although often researchers omit to reflect on what caused to system to collect some failures (Nissim et al., 2017), they are an ideal ground for sharing negative results and encourage reflections on "what did not work", an excellent opportunity to carry out a comparative error analysis and search for patterns that may, in turn, suggest improvements in both the dataset and the systems.

Here we analyze the case of the Hate Speech Detection (HaSpeeDe) task (Bosco et al., 2018) presented at EVALITA 2018, the Evaluation Campaign for NLP and Speech Tools for Italian (Caselli et al., 2018). HS detection is a really complex task, starting from the definition of the notion on which it is centered. Considering the growing attention it is gaining, see e.g. the variety of resources and tasks for HS developed in the last few years, we believe that error analysis could be especially interesting and useful for this case, as well as in other tasks where the outcome of systems meaningfully depends on resources exploited for training and testing.

The paper outlines the background and motivations behind this research (Section 2), describes the sub-task on which the study is based (Section 3), reports on the error analysis process (Section 4) and discusses its results (Section 5), and presents some conclusive remarks (Section 6).

2 Background and Motivations

There are several issues connected to the identification of HS: its juridical definition, the subjectivity of its perception, the need to remove potentially illegal content from the web without unjustly re-

moving legal content, and a list of linguistic phenomena that partly overlap to HS but need to be kept apart.

Many works have recently contributed to the field by releasing novel annotated resources or presenting automated classifiers. Two reviews on HS detection were recently published by Schmidt and Wiegand (2017) and Fortuna and Nunes (2018). Since 2016, shared tasks on the detection of HS or related phenomena (such as abusive language or misogyny) have been organized, effectively enhancing advancements in resource building and system development. These include HatEval at SemEval 2019 (Basile et al., 2019), AMI at IberEval 2018 (Fersini et al., 2018), HaSpeeDe at EVALITA 2018 (Bosco et al., 2018) and more. Nevertheless, the growing interest in HS detection suggests that the task is far from being solved: to improve quality and interoperability of resources, to design suitable annotation schemes and to reduce biases in the annotation is still as needed as it is to work on system engineering. Establishing standards and good practices in error analysis can enhance these processes and push towards the development of effective classifiers for HS.

While academic literature is rich with works on human annotation and evaluation metrics, it is not as easy to find works dedicated to error analysis of automated classification systems. This is rather more often found as a section of papers describing a system (see, e.g., (Mohammad et al., 2018)). This section, however, is not always present. To examine the errors made by a system, classify them and search for linguistic patterns appear to be a somewhat undervalued job, especially when the system had an overall good performance.Yet, it is crucial to understand why a system proved to be a weak solution to certain instances of a problem, even while being excellent for other instances.

In the context of COLING 2018, error analysis emerged as one of the most relevant features to be addressed in NLP research[1]. This attention to error analysis encouraged authors to submit papers with a dedicated section, with Yang et al. (2018) winning the award for the best error analysis, and is a step towards establishing good practices in the NLP community.

In the wake of this awareness, we apply linguistic insights to one of the annotated corpora used within the HaSpeeDe shared task, namely the HaSpeeDe-TW sub-task dataset (described in Section 3). Characteristics of this dataset make it ideal for our purpose: each tweet is connected to a target and is annotated not only for the presence of HS but for four other parameters. If a comparative analysis of two corpora presenting different textual genres (HaSpeeDe-TW and HaSpeeDe-FB) might have offered interesting perspectives, the lack of such characteristic in the FB dataset prevents a thorough comparison. Furthermore, among the in-domain HaSpeeDe sub-tasks, HaSpeeDe-TW is the one where systems achieved the lower F_1-scores, providing thus more material for our analysis.

3 HaSpeeDe-TW at EVALITA 2018: A Brief Overview

While a description of the HaSpeeDe task as a whole has been provided in the organizers' overview (Bosco et al., 2018), here we focus on HaSpeeDe-TW, one of the three sub-tasks into which the competition was structured[2]. The subtask consisted in a binary classification of hateful *vs* non-hateful tweets. Training set and test set contain 3,000 and 1,000 tweets respectively, labeled with *1* or *0* for the presence of HS, and with a distribution, in both sets, of around $1/3$ hateful against $2/3$ non-hateful tweets. Data are drawn from an already existing HS corpus (Poletto et al., 2017), whose original annotation scheme was simplified for the purposes of the task (see Section 4).

Nine teams participated in the task, submitting fifteen runs. The five best scores, submitted by the teams ItaliaNLP (whose runs ranked 1st and 2nd) (Cimino and De Mattei, 2018), RuG (Bai et al., 2018), InriaFBK (Corazza et al., 2018) and sbMMP (von Grünigen et al., 2018), ranged from 0.7993 to 0.7809 in terms of macro-averaged F_1-score[3]. They applied both classical machine learning approaches, Linear Support Vector Machine in particular (ItaliaNLP, RuG) and more recent deep learning algorithms, such as Convolutional Neural Networks (sbMMP) or Bi-LSTMs (ItaliaNLP, who adopted a multi-task learning approach ex-

[1] https://coling2018.org/error-analysis-in-research-and-writing/.

[2] The other two being HaSpeeDe-FB, where Facebook data were used both for training and testing the systems, and Cross-HaSpeeDe, further subdivided into Cross-HaSpeeDe-FB and Cross-HaSpeeDe-TW, where systems were trained using Facebook data and tested against Twitter data in the former, and the opposite in the latter.

[3] All official ranks are available here: https://goo.gl/xPyPRW.

ploiting the SENTIPOLC 2016 (Barbieri et al., 2016) dataset as well). Learning architectures resorted to both surface features such as word and character n-grams (RuG) and linguistic information such as Part of Speech (ItaliaNLP).

In the next section, we provide a description of the errors collected from these best five runs as put in relation with the specific factors we chose to analyze in this study, encompassing and merging qualitative and quantitative observations. Our analysis is strictly based on the results provided by those systems. An analysis focused on the features of the systems that determined the errors is unfortunately beyond the scope of this work, as in HaSpeeDe participants were only requested to provide the results after training their systems.

4 Error Analysis

Error analysis can be used in between runs to improve results or test different feature settings. With the aim of weaving a broader reflection on the especially hard linguistic patterns within a HS detection task, here it is performed *a posteriori* and on the aggregated results of five systems on the HaSpeeDe-TW test set (1,000 tweets). We focus on the answers given by the majority of the five best systems because we believe they provide a faithful representation of the errors without the noise due to the presence of the worst runs.

The test set was composed of 32.4% of hateful tweets and 67.6% non-hateful tweets. As the first step of our analysis, we compared the gold label assigned to each tweet in the test set with the one attributed by the majority of the five runs considered for the task. An error was considered to occur when the label assigned by the majority of the systems was different from the gold label. If we extend our analysis to all the fifteen submitted runs, 156 out of 1,000 tweets have been misclassified by the majority of them. However, this number increases to 172 if only the five best runs are taken into account.

Regardless of the correct label, agreement among the five best runs is higher than that among all runs and among any other set of runs: those systems which have best modeled the phenomenon on the data provided appear to have made similar mistakes. This supports our hypothesis that errors mostly depend on data-dependent features rather than on systems, which are all different in approach and feature setting.

Even though only the annotation concerning the presence of HS was distributed to the teams, the corpus from which the training and test set of HaSpeeDe-TW were extracted was provided with additional labels (Poletto et al., 2017; Sanguinetti et al., 2018). These labels (see Table 1) were meant to mark the user's intention to be aggressive (*aggressiveness*), the potentially hurtful effect of a tweet (*offensiveness*), the use of ironic devices to possibly mitigate a hateful message (*irony*), and whether the tweet contains any implicit or explicit reference to negative beliefs about the targeted group (*stereotype*).

label	**values**
aggressiveness	no, weak, strong
offensiveness	no, weak, strong
irony	yes, no
stereotype	yes, no

Table 1: The original annotation scheme of the HS corpus that was (partially) used in HaSpeeDe-TW.

These labels were conceived with the aim of identifying some particular aspects that may intersect HS but occur independently. As a matter of fact, hateful contents towards a given target might be expressed using aggressive tones or offensive/stereotypical slurs, but also in much subtler forms. At the same time, aggressive or offensive content, though addressed to a potential HS target, does not necessarily imply the presence of HS. Our assumption while carrying out this study was that such close, but at times misleading, relation between HS on one side and these phenomena on the other could be considered a source of error for the automatic systems.

In addition, other aspects of both linguistic and extra-linguistic nature were taken into account, so as to complement the analysis. We thus considered the tweets *targets*, i.e. Roma, immigrants and Muslims (also an information available from the original HS corpus). Finally, we selected three features that are typical of computer-mediated communication and social platforms such as Twitter, in particular, the presence of *links*, *multi-word hashtags*, and the use of *capitalized words*.

As for the method adopted, the percentage of errors for the gold positives and the gold negatives in the whole test set was calculated. First, the rates were calculated considering the two labels - hateful and non-hateful - separately, in order to bal-

ance their different distribution in the test set; then the results were halved to represent the whole corpus in percentage and to maintain the proportion between the results of the tags. All the percentages correlating two different tags were calculated this way, so that the results could be easily compared. The percentages of mistakes for each label of the categories were determined and compared to the general result to understand whether they influenced it positively or negatively. Table 2 summarizes the results for each label showing the distribution of the false negatives (FN), false positives (FP), true positives (TP) and true negatives (TN). The error percentages higher than the general result are in bold font.

5 Results and Discussion

In order to find some answers to our research questions and evidence of the influence of the annotated features on the systems' results, we provide in this section an analysis driven by the categories we described in the previous section.

Aggressiveness and Offensiveness. The different degrees of aggressiveness did not affect the systems recall, but we measured more FPs when weak or strong aggressiveness is involved (more than thrice as many as in the overall results when strong aggressiveness is present).
Offensiveness seems to hold a similar but heavier influence on performance, causing better recall but worse precision: FPs are more than doubled when strong offensiveness is present.

The presence of offensiveness is often associated to slurs or vulgar terms: these are not a consistent presence in the dataset (the most vulgar tweets are probably quickly removed by the platform), and mostly appear in tweets classified as HS. However, about half of the non-hateful tweets containing offensive words were wrongly classified as hateful, proving that offensiveness can be misleading for systems. In these cases, a lexicon-based approach can fail, while attention to the context could be crucial: in the most common instances of false positives, in fact, offensive words did not refer to the targets.

HS Targets. Analyzing the three targets of HS allowed us understanding how the systems reacted to different ways of expressing hate.

Most of the errors were caused by the target *Roma*: few hateful tweets were recognized, and FNs are more than 30%. Results for the target *Immigrants* are similar to the overall performance, only with a slightly higher number of FPs. The target *Muslims* caused a low number of FNs but almost twice as many FPs as in the general performance.

The systems seem to struggle to recognize hateful content against Roma: this may be caused by an imbalance in the test set (only 6.3% of tweets with the target *Roma* are labelled as HS, while the targets *Immigrants* and *Muslims* have 12.6% and 13.4% of hateful tweets respectively) or by biases in the annotation.

The poor results achieved in classifying messages with target *Roma* can also be explained by the subtler ways of expressing HS when this target is involved, more heavily based on stereotypes than it happens with the other targets. The hate against the other two targets, in particular Muslims, was instead very explicit. See the following examples extracted from the test set.

> 2235. *Roma, colpisce una pecora con il pallone: bambino rom accecato da un pastore https://t.co/KsSAS3fUx9 @ilmessaggeroit HA DIFESO I SUOI AVERI!*[4] [FN, strong aggressiveness, target: Roma]

> 4749. *@Corriere Uccidere gli islamici, prima di tutto.*[5] [TP, strong aggressiveness, target: religion]

Other features. Some other features were considered in our analysis. The presence of *stereotype* was more frequent in hateful tweets, which caused a slight increase in FPs; conversely, cases of HS without stereotype posed no issues to the systems. Moreover, as expected, the presence of *irony* slightly increased the errors rate both in hateful and non-hateful tweets.

The presence of Twitter's linguistic devices also negatively influenced the results, probably because of the difficulty encountered by systems when some semantic content assumes non-standard forms, e.g. *links*, *multi-word hashtags* and *capitalized words*.

URLs frequently occur in the data, but mostly in non-hateful tweets (although this may be a peculiarity of this dataset). Systems appear to have

[4] "Rome, Roma child hits a sheep with a ball: blinded by a shepherd https://t.co/KsSAS3fUx9 @ilmessaggeroit HE DEFENDED HIS PROPERTY!"

[5] "@Corriere Kill the Muslims, first of all."

	FN	FP	TP	TN	Gold HS	Gold Not-HS
general	15%	6%	35%	44%	32.3%	67.7%
no aggressiveness	15%	4%	35%	46%	13.5%	56.8%
weak aggressiveness	15%	**10%**	35%	40%	11.2%	10.1%
strong aggressiveness	15%	**19%**	35%	31%	7.6%	0.8%
no offensiveness	**20%**	5%	30%	45%	10.9%	60%
weak offensiveness	13%	**11%**	37%	39%	14.6%	4.9%
strong offensiveness	12%	**16%**	38%	34%	6.8%	2.8%
no irony	15%	5%	35%	45%	27.8%	59%
yes irony	**18%**	**9%**	32%	41%	4.5%	8.7%
no stereotype	15%	5%	35%	45%	11.6%	49.7%
yes stereotype	15%	**8%**	35%	42%	20.7%	18%
Immigrants	15%	**9%**	35%	41%	12.6%	22.4%
Muslims	8%	**11%**	42%	39%	13.4%	12.2%
Roma	**31%**	1%	19%	49%	6.3%	33.1%
no link	11%	**13%**	37%	39%	25.4%	24.4%
yes link	**29%**	1%	21%	49%	7%	43.2%
multi hashtags	**23%**	**8%**	27%	42%	3%	1.9%
no capitalized words	15%	5%	35%	45%	29.1%	64.1%
yes capitalized words	14%	**9%**	36%	41%	3.3%	3.5%

Table 2: Percentage of correct (TPs and TNs) and erroneous (FPs and FNs) results in relation to the features considered in the analysis, along with the actual distribution of these features in the test set.

troubles recognizing hateful tweets that contain URLs (errors increased by 14%). Conversely, the absence of URLs caused an increase in FPs. This feature is unlikely to be directly connected to hateful language: we rather believe that it could somehow affect predictions regardless of the actual content.

Also multi-word hashtags influenced results, especially for hateful content: their presence increased FNs by 8%. The reason for this kind of error might lie in the fact that our dataset contains some cases where the crucial element in a hateful tweet is precisely the hashtag, as in the example below:

> 2149. *Quando vedremo lo stessa tema portato in piazza con la stessa forza e determinazione? Mai credo. #stopislam* [6] *https://t.co/dDYLZB1BlJ* [multi-word hashtag, FN]

The text in this tweet is not hateful, but an element of hatred is conveyed by the hashtag "#stopislam".
The ability to separate the multi-word hashtags into the words composing them would improve the performances of the systems. The tweets with a multi-word hashtag clarifying the text would have a better chance of being correctly identified.

Finally, some capitalized words have been found in the data set, mostly in hateful tweets, which again caused an increase in FPs. Despite their small number, we noticed that, in non-hateful tweets, a higher percentage of capitalized words are named entities (nouns of places, people, newspapers, etc.), while in hateful tweets capitalized words are more often used to intensify opinions or feelings.

Among all the features taken into account, offensiveness seems to have affected the performance in various ways: its absence led systems to classify as non-hateful tweets that are indeed hateful, while its presence caused the inverse error. A possible explanation for this is that, as shown in Sanguinetti et al. (2018), offensiveness does not correlate with HS even though it can be one of its features. The systems might have taken offensive terms as indicators for HS, as also humans tend to do (see for example Bohra et al. (2018)), but this is a false assumption that systems should be trained to avoid. Aggressiveness also caused a certain degree of errors, but only affecting precision.

[6] "When will we see people fighting for the same issue with the same strength and determination? Never, I believe."

6 Lessons Learned and Conclusion

This paper presents a detailed error analysis of the results obtained within the context of a shared task for HS detection. In our study, we took into account two types of data: content information, provided by gold standard labels assigned to each tweet; and metadata information, namely the presence of URLs, hashtags and capitalized words. Results prove the importance of considering other categories related to that on which the task was centered.

The analysis of performances in relation to URLs poses a controversial result. There are two reasons why tweets collected via Twitter's API may contain a URL: the tweet may have been cut off and a URL automatically generated as a link to the complete tweet, or the URL may be part of the original tweet and lead to an external page. In both cases, unless the URL is followed, the tweet is likely to be harder to understand compared to a tweet that contains no URL. This may cause lower agreement among human judges, and it is a very complicated issue for automated systems to deal with, especially when the meaning of the tweet is unintelligible without first opening the URL. Tweets containing URLs are, for the time being, less reliable as training data and pose a tougher challenge for Sentiment Analysis tasks at large; we encourage an effort towards solving this issue.

As for capitalized words, future work may include investigating how they affect human annotation, as some judges may show a bias towards associating capitalized words to HS or other categories. Furthermore, improvements may come from considering the PoS tags of such words, or the number of consecutive capitalized words.

Multi-word hashtags as well need to be treated with care, as they may affect and even overturn the meaning of the whole tweet. Yet, it happens that a hashtag might require syntactic, semantic and world-knowledge processing in order to be fully understood: for example, by comparing the phrase "stop Islam" with, e.g., "stop harassment", we can see that the word "stop" is not necessarily negative, and it becomes so only because it is followed by the name of a religion whose members are, nowadays and in Western society, particularly subject to discrimination.

Overall, our analysis suggests that systems failures are motivated by the difficulty in dealing with cases where HS is less directly expressed and pave the way for future work on, e.g., the development of tools that perform a more careful analysis of the text.

Acknowledgments

The work of C. Bosco and M. Sanguinetti is partially funded by Progetto di Ateneo/CSP 2016 (*Immigrants, Hate and Prejudice in Social Media*, S1618_L2_BOSC_01), while that of F. Poletto is funded by Fondazione Giovanni Goria and Fondazione CRT (*Talenti della Società Civile 2018*).

References

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. RuG @ EVALITA 2018: Hate Speech Detection In Italian Social Media. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the Evalita 2016 SENTIment POLarity Classification Task. In *Proceedings of the Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*. CEUR.org.

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.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media*, pages 36–41.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 hate speech detection task. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Tommaso Caselli, Nicole Novielli, Viviana Patti, and Paolo Rosso. 2018. EVALITA 2018: Overview of the 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Andrea Cimino and Lorenzo De Mattei. 2018. Multi-task Learning in Deep Neural Networks for Hate Speech Detection in Facebook and Twitter. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Michele Corazza, Stefano Menini, Pinar Arslan, Rachele Sprugnoli, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Comparing Different Supervised Approaches to Hate Speech Detection. In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, pages 214–228. CEUR-WS.org.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17.

Malvina Nissim, Lasha Abzianidze, Kilian Evang, Rob van der Goot, Hessel Haagsma, Barbara Plank, and Martijn Wieling. 2017. Sharing is caring: The future of shared tasks. *Computational Linguistics*, 43(4):897–904.

Fabio Poletto, Marco Stranisci, Manuela Sanguinetti, Viviana Patti, and Cristina Bosco. 2017. Hate Speech Annotation: Analysis of an Italian Twitter Corpus. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*. CEUR.org.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics.

Dirk von Grünigen, Ralf Grubenmann, Fernando Benites, Pius Von Däniken, and Mark Cieliebak. 2018. spMMMP at GermEval 2018 Shared Task: Classification of Offensive Content in Tweets using Convolutional Neural Networks and Gated Recurrent Units. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*.

Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.