Acquiring a Formality-Informed Lexical Resource for Style Analysis

Elisabeth Eder[1], **Ulrike Krieg-Holz**[1], **Udo Hahn**[2]
[1] Institut für Germanistik, Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
[2] Jena University Language & Information Engineering (JULIE) Lab,
Friedrich-Schiller-Universität Jena, Jena, Germany
{elisabeth.eder, ulrike.krieg-holz}@aau.at, udo.hahn@uni-jena.de

Abstract

To track different levels of formality in written discourse, we introduce a novel type of lexicon for the German language, with entries ordered by their degree of (in)formality. We start with a set of words extracted from traditional lexicographic resources, extend it by sentence-based similarity computations, and let crowdworkers assess the enlarged set of lexical items on a continuous informal-formal scale as a gold standard for evaluation. We submit this lexicon to an intrinsic evaluation related to the best regression models and their effect on predicting formality scores and complement our investigation by an extrinsic evaluation of formality on a German-language email corpus.

1 Introduction

The computational treatment of style in verbal communication has long been dominated by application concerns, e.g., the identification or profiling of authors in forensic linguistics (Ding et al., 2019) or the recognition of plagiarism (Alzahrani et al., 2012). This research was conducted assuming that simple lexico-statistic patterns identified by stylometric computations were sufficient to solve authorship and plagiarism assignment problems.

Despite their undisputed success in those limited fields, these studies scratched only the surface of the notion of 'style' as discussed in linguistic pragmatics (Hickey, 1993). From the many ways 'style' can be approached from a pragmatics perspective, we here focus on its inherent formality dimension, i.e., the distinction between *formal* (standard) and *informal* (colloquial) language use (for a survey, cf. Heylighen and Dewaele (1999)), with further extensions directed at the higher level of formal (e.g., elevated style) and the lower level of informal (e.g., vulgar) phrasing. Such distinctions of formality levels are crucial for the appropriateness of verbal expressions in a given discourse context.

In order to track different levels of formality in written communication, we introduce a novel type of lexicon for the German language, with entries ordered by their degree of (in)formality.[1] We start with a set of words extracted from traditional lexicographic resources, extend it by sentence-based similarity computations, and let crowdworkers assess the enlarged set of lexical items on a continuous informal-formal scale. This workflow is described in Section 3. The resulting lexicon comprising words with their respective formality scores subsequently serves as a gold standard for evaluation. In Section 4, we submit this lexicon to an intrinsic evaluation related to the best regression models and their effect on predicting formality scores, and complement our investigation by an extrinsic evaluation of formality on a German-language email corpus in Section 5.

2 Related Work

The relevance of 'style' for NLP is obvious for language output-focused core applications such as language generation (Sheikha and Inkpen, 2011; Dethlefs et al., 2014; Ficler and Goldberg, 2017), machine translation (Niu et al., 2018; Prabhumoye et al., 2018) or proper phrasing in argumentation (El Baff et al., 2020). Quite recently, the notion of 'formality style transfer' has received increasing attention, which captures the idea to generate a formal sentence given an informal one (*et vice versa*), while preserving its meaning (Shen et al., 2017; Fu et al., 2018; Rao and Tetreault, 2018; Li et al., 2018; Prabhumoye et al., 2018; Yang et al., 2018; Lample et al., 2019; Gong et al., 2019; Dai et al., 2019; Wu et al., 2019; John et al., 2019; Luo et al., 2019; Wang et al., 2019; Shang et al., 2019; Wang et al., 2020; Zhang et al., 2020; Yi et al., 2021).

[1] The lexicon is available at https://github.com/ee-2/I-ForGer.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2028–2041
April 19 - 23, 2021.

Many efforts to cope with language style have been spent, however, in application niches, such as author identification or plagiarism detection. Most of the methodological contributions developed in this forensic branch are summarized under the label of *stylometrics* and have recently found their way into NLP analytics to unveil deception (Potthast et al., 2018; Pascucci et al., 2020a) or linguistic aggression (Harpalani et al., 2011; Nogueira dos Santos et al., 2018; Pascucci et al., 2020b).

The computational analysis of style according to stylometric principles, from its inception, is closely linked with *lexical frequency* counts. Typically, mostly function words (such as articles, pronouns, conjunctions, contractions, common abbreviations, hedging terms, also including punctuation marks) are assembled in small-sized dictionaries, together, if at all, with only a few content words (domain-specific nouns, verbs, adjectives). The frequency distributions resulting from counting these dictionary entries at the document or corpus level are already very beneficial for successfully dealing with disputed authorship problems (mostly for literary texts, but also for the detection of spam, fake news, or other kinds of toxic language) or uncovering plagiarism (mostly in scientific or news documents). Similar in spirit, word and sentence length criteria originating from readability metrics (Flesch-Kincaid, etc.) and several measures of vocabulary richness (e.g., type-token ratios, Yule's K and Burrow's Δ) were also incorporated into stylometric toolkits (Eder et al., 2016).

As a simple extension from these uni-grams, lexical or pseudo-lexical character *n-grams* (bi- or tri-grams, mostly) were determined and counted, as well. Slightly extending this (pseudo-)lexically focused approach by syntactic information, part-of-speech n-grams (or part-of-speech frequencies) were also considered to trace the human 'stylome', although lexical factors were found to be more relevant for style analysis than syntactic (POS sequence) patterns (van Halteren et al., 2005).

Simple frequency metrics have increasingly been complemented by various forms of lexical *association measures* (such as information gain, mutual information), and more sophisticated probabilistic models (principal component analysis (PCA), latent semantic analysis (LSA), or other types of topic models). Comprehensive lists of criteria and metrics are provided by Sheikha and Inkpen (2010); Neal et al. (2017); Ding et al. (2019).

We claim that despite their relevance for applications, such as authorship attribution and plagiarism detection, these mechanisms merely serve as easy to trace proxies for characterizing linguistic style. In our work, we will have a closer look at the style-marking semantic connotation of single lexical items as explicit carriers of linguistic formality as an important facet of language style.

A milestone for the formal definition of formality was set up by the pioneering work of Heylighen and Dewaele (1999) who defined the $\mathcal{F}$*-score*—close in spirit with the simple *lexico-statistic frequency* metrics from stylometry—as the percentage difference between deictic (article, pronouns, etc.) and non-deictic parts of speech (nouns, adjectives, etc.) in a document ($\mathcal{F}$ ranges between 0 and 100, with higher $\mathcal{F}$ indicating higher formality).[2] This document-level perspective was adapted by Lahiri et al. (2011) to sentence-level formality analysis.

A complementary *lexical* dimension for the formalization of formality was introduced by Brooke et al. (2010). They define the *formality score* for a word as a real number value in the range 1 to -1, with 1 representing an extremely formal word and -1 an extremely informal one, and assign a formality score to each lexical item based on standard word length, morphology-based features, lexical distribution criteria or association methods (LSA). Our work adheres to their way formality is scored in a formality lexicon and manually supplied seed sets are used (as starters), but differs markedly whether the lexicon is considered as a static (Brooke et al., 2010) or a dynamic resource (as we do; in a later study, Brooke and Hirst (2014) proposed a dynamic acquisition method, as well, by assigning a continuous formality score to single words based on their co-occurrence frequency with a hand-picked seed set of formal, neutral and informal words), and the way how semantic similarity is computed (LSA vs. embeddings). Further, we do not induce formality levels for a near-synonym task automatically but rather crowdsource nuances of formality for a relationally unrestricted lexical inventory from human raters.

Pavlick and Tetreault (2016) proposed a model of formality based on an empirical analysis of human formality perceptions. They apply their approach to analyze language use in online debate forums for multiple genres (news, blogs, emails,

[2]The $\mathcal{F}$*(ormality)*-score must not be confused with the F-score as a measure relating precision and recall.

and community question answering sites). Formality assessments are solicited via Amazon Turk (following the protocol established by Lahiri (2015)) using a 7-point Likert scale, with labels ranging from −3 (Very Informal) to 3 (Very Formal). A ridge regression classifier uses 11 different feature groups—five rarely used ones (among them WORD2VEC embeddings (Mikolov et al., 2013), parse trees, dependency tuples, and named entities) and six much more common ones (among them lower/upper casing, punctuation, readability scores, POS tags, and length-normalized formality and subjectivity scores)—to determine the formality level of sentences. Cross-genre analysis reveals that n-grams and word embeddings perform the best among all tested features (they achieve over 80% of the performance of the full classifier in all cases). This work comes closest to our approach, yet with differences in the way formality is assessed (Likert scales vs. best-worst scaling) and lexicon building is dealt with. Pavlick and Tetreault (2016) employ an acquisition method to score the formality of unseen phrases along the formal-casual dimension from scratch, as described in earlier work by Pavlick and Nenkova (2015) who use a log ratio metric based on the occurrence of phrases in various style-tagged corpora, in contrast to the embedding-based similarity model we propose.

Earlier computational models for detecting formality were proposed by Sheikha and Inkpen (2010); Peterson et al. (2011); Mosquera and Moreda (2012). The first two perform a binary classification only into formal *vs.* informal utterances, the third model classifies into four levels of (in)formality, and all of them operate at the document (as opposed to sentence) level.

3 Building a Formality-Informed Lexicon

3.1 Getting Started with VULGER

Previous work on computational lexicons (and lexicon acquisition) incorporating formality information focuses exclusively on the English language (Brooke et al., 2010; Brooke and Hirst, 2014; Pavlick and Nenkova, 2015; Pavlick and Tetreault, 2016). For German, VULGER (Eder et al., 2019)[3] constitutes a lexical resource that can be reused for such purposes to some degree. It comprises 3,300 German words scored by vulgarity/neutrality within a range of −1 (most vulgar) to +1 (most neutral). Accordingly, it covers the lower half of the formality spectrum quite well but completely lacks its upper half (formal up to elevated language). This study attempts to fill this gap by introducing I-FORGER, a comprehensive lexicon for *I*nformal and *For*mal *Ger*man. To acquire a lexicon covering the formal spectrum as well, we gathered formality-marked lexical entries in several ways as described in the following subsections.[4]

3.2 Input from Lexicographic Resources

As a first lexical acquisition step, we gathered lexical items from existing lexicographic resources based on their *manually* assigned categorical (in)formality tags:

Swear Words. As there is an overlap between swear words and vulgar lexicalizations, we used 500 lexical items randomly chosen from three German swear word lists[5] to feed the lower end of formality in I-FORGER.

Colloquial Items. In addition, we extracted 500 arbitrary terms marked as 'colloquial' ('ugs.' or 'umgangssprachlich,' in German) from the German slice of WIKTIONARY[6] and the German OPENTHESAURUS[7] supposed to range somewhere between vulgar and neutral on our scale.

Elevated Items. To extend the scale to the upper levels of linguistic formality, we also picked lexical items marked as 'elevated' ('geh.' or 'gehoben,' in German) from OPENTHESAURUS and WIKTIONARY yielding 1,000 additional terms (for the sake of balancing informal and formal entries in that phase). The reuse of manually curated lexicon resources (as seeds) thus follows the approach proposed by Brooke et al. (2010).

3.3 Lexicon Extension via Sentence Similarity

Given the intrinsic limitations of any manually curated lexicon resource, in the next step, we augmented I-FORGER by *automatic* means. We here suggest harvesting lexical candidates potentially carrying formality information from semantically

[3] https://github.com/ee-2/VulGer

[4] For basic NLP processing routines, we used SPACY (Honnibal et al., 2020) and FLAIR (Akbik et al., 2018, 2019).

[5] Retrieved from http://www.hyperhero.com/de/insults.htm, http://www.insult.wiki/wiki/Schimpfwort-Liste and https://www.schimpfwoerter.de on April 24, 2020.

[6] https://de.wiktionary.org

[7] https://www.openthesaurus.de

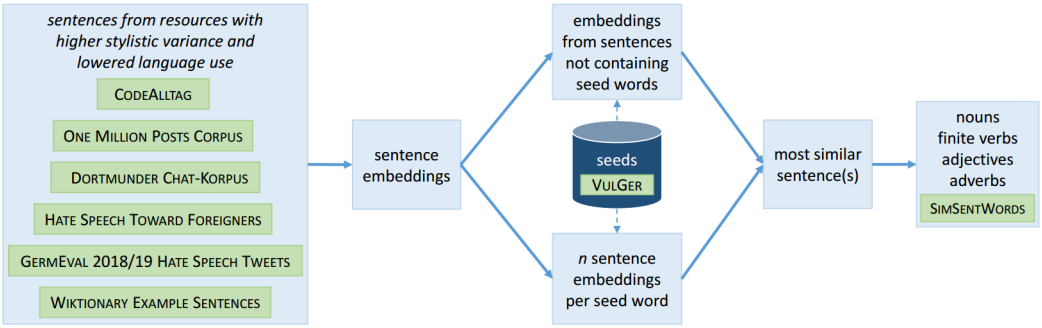

Figure 1: Generic language-independent workflow for gathering words for formality scoring approaches utilizing similar sentences (in blue) and its instantiation for our use case to acquire SIMSENTWORDS (in green)

similar sentences. This proposal goes beyond the standard way to utilize word embeddings in order to find close semantic neighbors based on the distributional hypothesis (see, e.g., Tulkens et al. (2016); Wiegand et al. (2018a) for detecting abusive lexicalizations this way). Rather than only discovering semantically related *words*, we extended our scope to semantically similar *sentences* to identify other relevant lexical candidates in the mined sentences, like an adjective modifying an offensive noun or other vulgar, yet otherwise unrelated, words in a vulgar word's context. On the flip side, this method admittedly gathers a considerable amount of noise (cf. Section 4 for a scoring approach to account for this problem).

3.3.1 Sentence Embeddings

As is well-known, BERT (Devlin et al., 2019) reaches new state-of-the-art results for various NLP problems, including semantic similarity tasks. However, finding semantically similar sentences close in vector space with BERT is computationally expensive. As a cure, Reimers and Gurevych (2019) introduced SENTENCE-BERT (SBERT), which modifies the pre-trained BERT network using siamese and triplet networks and produces semantically meaningful sentence embeddings that can be compared employing standard cosine similarity.

3.3.2 Sentence Similarity

To obtain candidate sentences for similarity computation for the German language, we employed a wide range of corpora. Our choices were guided by the requirements that these corpora should possess a high stylistic variance and contain vocabulary from the lower language register, too. We came up with:

- CODE ALLTAG[8] (Eder et al., 2020) comprising roughly 1,5M German-language emails,
- ONE MILLION POSTS CORPUS[9] (Schabus et al., 2017) containing about 1M user comments on news articles from the Austrian daily broadsheet newspaper DER STANDARD,
- DORTMUNDER CHAT KORPUS[10] (Beißwenger, 2013), with more than 140,000 German-language chats,
- HATE SPEECH TOWARDS FOREIGNERS[11] (Bretschneider and Peters, 2017), with about 6,000 posts and comments on German anti-foreign FACEBOOK pages,
- GERMEVAL 2018/19, collected for the task of identifying offensive language[12] (Wiegand et al., 2018b; Struß et al., 2019), including roughly 15,000 German-language tweets.

Using VULGER as a seed lexicon, we extracted sentences from these corpora by separating those containing VULGER entries from those that did not contain any VULGER item.[13] To further enlarge the number of sentences for each seed item, we also gathered sentences given as examples on the WIKTIONARY pages for the entries included in VULGER. From the resulting pool of sentences with seed words, we collected up to six sentences

[8] https://github.com/codealltag
[9] https://ofai.github.io/million-post-corpus
[10] https://www.uni-due.de/germanistik/chatkorpus
[11] http://www.ub-web.de/research/index.html
[12] https://projects.fzai.h-da.de/iggsa/data-2019
[13] We only took 100,000 randomly chosen sentences from CODE ALLTAG and the ONE MILLION POSTS CORPUS for performance reasons.

per word. We chose them randomly but tried to take one sentence from each of the six resources to keep some balance, both formality-wise as well as genre-wise.

These sentences served as seeds for the computation of similar sentences. Like the remaining sentences not containing any seed words, they were embedded with SENTENCE TRANSFORMERS (Reimers and Gurevych, 2019) using the multilingual model supporting German (Reimers and Gurevych, 2020). Then, for all seed sentence embeddings, we calculated the most similar sentence in the remaining sentence embeddings using cosine distance (the acquisition step proper). From these most similar sentences, we gathered lemmatized nouns, finite verbs, adjectives, and adverbs, omitting named entities. An overview of the entire acquisition procedure is depicted in Figure 1.

From the resulting word list, we randomly chose 1,000 items (denoted SIMSENTWORDS, in the following) to evaluate the regression approach and the acquisition strategy of automatically gathering new words to score. As we also wanted to measure the acquisition noise, we further divided the words into 500 items manually cleansed from spelling mistakes, etc. (SIMSENTWORDS$_{cleansed}$), and left 500 as-is (SIMSENTWORDS$_{noisy}$).

3.4 I-FORGER at a Glance

Putting these pieces together, I-FORGER, the final lexicon, comprises 3,000 words, in total, with three major divisions: 1,000 terms from elevated language usage, 1,000 words, with swearwords and colloquial items joined, presumably linked to the lowered stylistic inventory, and 1,000 words that should rather occur at the lower end of our informality scale, but potentially include words from all stylistic levels (see Table 1).

Resource	# Lexical Items
ELEVATEDWORDS	1,000
SWEARWORDS	500
COLLOQUIALWORDS	500
SIMSENTWORDS	1,000
SIMSENTWORDS$_{cleansed}$	500
SIMSENTWORDS$_{noisy}$	500
Total	3,000

Table 1: Contributions from various resources for the I-FORGER lexicon

3.5 Human Assessment of I-FORGER

To establish a gold standard for subsequent evaluation, we gathered human formality assessments. For that, I-FORGER was annotated with Best-Worst-Scaling (BWS), a method that delivers high-quality annotations with only a relatively small number of annotation steps compared to standard point-interval based methods (e.g., Likert scales) for human assessment tasks. BWS also adheres to the principle that a "continuum of formality" (Heylighen and Dewaele, 1999) exists rather than n-ary categorical distinctions between formal and informal utterances (see also Lahiri et al. (2011); Brooke and Hirst (2014) for works based on degrees of formality).

BWS was introduced into NLP for emotion scaling by Kiritchenko and Mohammad (2016, 2017). Annotators are presented with n items at a time (an n-tuple, where $n > 1$, and typically $n = 4$). They then have to decide which item from the n-tuple is the *best* (highest in terms of the property of interest) and which is the *worst* (lowest in terms of the property of interest).

In our case, judges had to select the *most elevated* and the *most vulgar* terms per given n-tuple. We used the BWS tool[14] from Kiritchenko and Mohammad (2016, 2017) to generate 6,000 4-tuples for human assessment. Tuples were produced randomly under the premise that each term had to occur only once in eight different tuples and each tuple was unique.

For the annotation process proper, we used the crowdsourcing platform CLICKWORKER,[15] where we had each n-tuple assessed by five annotators (Kiritchenko and Mohammad (2016) showed that as few as 2-3 responses per tuple are sufficient to get reliable scores, at least for the assessment of sentiment.). In order to get real-valued scores from the BWS annotations, we applied COUNTS ANALYSIS (Orme, 2009)[16] and subtracted the percentage of times the term was chosen as worst from the percentage of times the term was chosen as best. Thus, we got scores between $+1$ (most formal) and -1 (most informal). We computed the split-half reliability[16] by randomly splitting the annotations of a tuple into two halves, calculating scores independently for these halves, and mea-

[14] http://www.saifmohammad.com/WebPages/BestWorst.html

[15] https://www.clickworker.de

[16] Again, we used the scripts from Kiritchenko and Mohammad (2016, 2017).

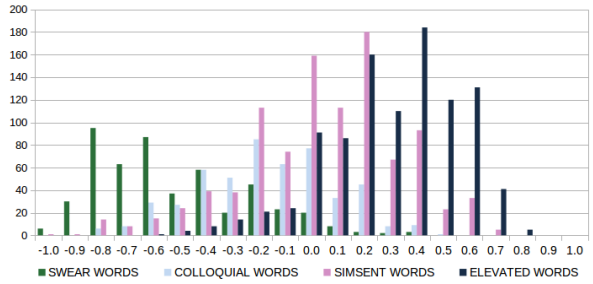

Figure 2: Distribution of scores per resource of I-FORGER

suring the correlation between the resulting two sets of scores. We got an average Spearman's ρ of 0.8954 $(+/-0.0030)$ over 100 trials.

Figure 2 displays the distribution of human assessed scores per resource for I-FORGER. While SWEARWORDS and, to a lesser degree, also COLLOQUIALWORDS are linked to lower scores, and ELEVATEDWORDS obtained higher scores, SIMSENTWORDS are found in the middle spreading on the entire scale of scores, also comprising a fair amount of words from the lower end of formality.

4 Intrinsic Evaluation of I-FORGER

Rather than increasing the size and thus the coverage of lexicons to improve performance on potential applications, we intend to score (unseen) words on the fly. Hence, we first evaluate the word scoring model (Section 4.1). Next, we assess the four main input streams of I-FORGER (Section 4.2) and the extension of the scale regarding formality levels (Section 4.3). Figure 3 illustrates the schematic workflow for our word scoring procedure, including (and marked in green) the three evaluation tasks.

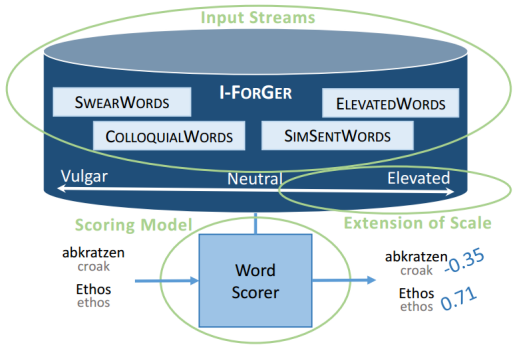

Figure 3: Overview of the word scoring workflow; parts to evaluate are marked in green

4.1 Regression Models for Word Scoring

We adopted various approaches using a seed lexicon, actually, the entries' word embeddings, as training data for regression models to automatically score new lexical items for their formality connotation (see, e.g., Li et al. (2017) and Buechel and Hahn (2018) for a similar scenario for automatic emotion induction).

As input features we decided for FASTTEXT word embeddings (Grave et al., 2018) with their own out-of-vocabulary (OOV) functionality. We found that they performed better than getting the OOV handling from BPEMB subword embeddings (Heinzerling and Strube, 2018), based on Byte Pair Encoding (BPE) (Sennrich et al., 2016), instead, or solely utilizing pure BPEMB embeddings.

We evaluated different regression models. Besides RIDGE REGRESSION,[17] which is linear regression with L_2 regularization during training, we also experimented with DENSIFIER (Rothe et al., 2016), which learns an orthogonal transformation of the embedding space, and a modified, more robust variant of the latter, DENSRAY (Dufter and Schütze, 2019).[18] We ran a feed-forward neural network with one hidden layer combined with the boosting algorithm AdaBoost.R2 (BOOSTED FFNN) as proposed by Du and Zhang (2016).[19] Further, we tested neural networks with more than one hidden layer, namely two hidden layers with 256, 128 units ($\text{NN}_{2Hidden}$) and three hidden layers with 256, 128 and 64 units ($\text{NN}_{3Hidden}$).[20]

Table 2 depicts that DENSIFIER and DENSRAY performed worse than all the others. Also, RIDGE REGRESSION yielded significantly lower results than the BOOSTED FFNN model. We found no difference between $\text{NN}_{2Hidden}$, $\text{NN}_{3Hidden}$ and BOOSTED FFNN since all three reached a strong Spearman's ρ of 0.77. As a higher number of hidden layers did not significantly improve results, we used BOOSTED FFNN for further processing.

[17] We used the SCIKIT-LEARN.ORG implementation with the default parameters.

[18] We used their code provided on https://github.com/pdufter/densray.

[19] We copied their code on https://github.com/StevenLOL/ialp2016_Shared_Task.

[20] We used KERAS in TENSORFLOW with the following hyperparameters: embedding/input layer with 0.2 and hidden layers with 0.5 dropout, MaxNorm weight constraint of 3, random normal weight initialization, ReLu activation, Adam optimizer, batch size of 32, mean squared error loss and 1,000 epochs with early stopping.

Model	Spearman's ρ
RIDGE REGRESSION	0.706*
DENSIFIER	0.632*
DENSRAY	0.621*
$NN_{2Hidden}$	0.773
$NN_{3Hidden}$	0.771
BOOSTED FFNN	0.773

Table 2: Averaged Spearman's ρ for different models (10-fold cross-validation on I-FORGER); statistically significant differences (using the two-sided Wilcoxon signed-rank test on Spearman's ρ) are marked with '*' for $p < 0.005$ with respect to BOOSTED FFNN

4.2 Assessment of Input Streams

Table 3 pinpoints the predictability of formality for a particular input stream of I-FORGER in a 10-fold cross-validation setting. Learning scores of COLLOQUIALWORDS and ELEVATEDWORDS seems harder than scoring SWEARWORDS and SIMSENTWORDS. The lower human agreement on choosing the most elevated item supports this finding for the upper half of the formality spectrum. The data also reveal that the regression model is somewhat prone to noise since original SIMSENTWORDS$_{noisy}$ achieved much lower results than curated SIMSENTWORDS$_{cleansed}$. However, this acquisition strategy seems to be a choice worth considering for scoring approaches.

Input Stream	Spearman's ρ
SWEARWORDS	0.593
COLLOQUIALWORDS	0.409
SIMSENTWORDS	0.672
SIMSENTWORDS$_{cleansed}$	0.732
SIMSENTWORDS$_{noisy}$	0.595
ELEVATEDWORDS	0.477
I-FORGER	0.773

Table 3: Spearman's ρ for BOOSTED FFNN on I-FORGER with results for different input streams (10-fold cross-validation)

4.3 Assessment of Formality Scale Extension

A comparison with VULGER suggests that scoring an extended range of linguistic styles is a more difficult task, since evaluating the BOOSTED FFNN model on VULGER achieved a higher Spearman's ρ of 0.827 (10-fold cross-validation) than on I-FORGER (see Table 3). Nevertheless, applying a model trained on I-FORGER to VULGER gave a Spearman's ρ of 0.678, which signals evidence that I-FORGER still captures the vulgar-neutral dimension despite being trained on an extended scale with fewer words (3,000 *vs.* 3,300). It also shows that the word scoring approach *per se* indeed yields reliable results on the informal-formal dimension.

5 Extrinsic Evaluation of I-FORGER

In order to gather evidence for the value of I-FORGER in combination with the word scoring approach within a realistic use case, we ran experiments with emails, which possess a higher stylistic variability than news concerning their formality spread (Pavlick and Tetreault, 2016). Other work related to the formality of emails is typically carried out in the context of communication behavior studies in enterprises, with a focus on determining social factors (social distance, relative power, and the weight of imposition) that affect the sender's choice of formality (Peterson et al., 2011) or on the affective dimension of email exchanges (Chhaya et al., 2018) in terms of the prediction of frustration of employees from email data.

5.1 Email Corpus and Formality Gold Standard

Again using BWS and the tools from Kiritchenko and Mohammad (2016, 2017) mentioned before, we manually scored 800 German emails from CODE ALLTAG$_{S+d}$, a specialized, metadata-rich subset of CODE ALLTAG (Eder et al., 2020), for their formality. 35 annotators had to select the *most formal* email and the *most informal* email from four emails per rating step. Altogether, we had 1,600 4-tuples assessed three times. We got an average

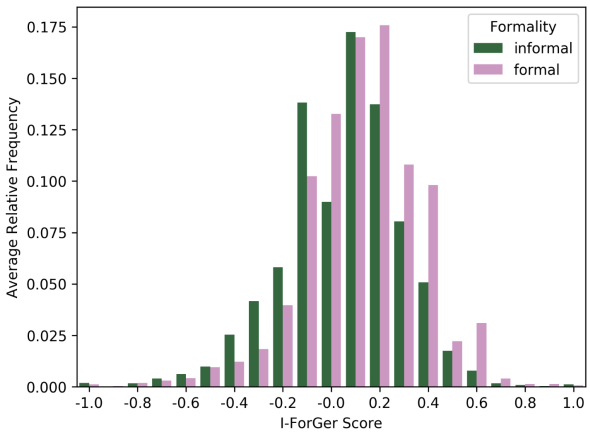

Figure 4: Distribution of I-FORGER scores for formal (with formality scores from 0 to +1) and informal emails (rated from −1 to 0)

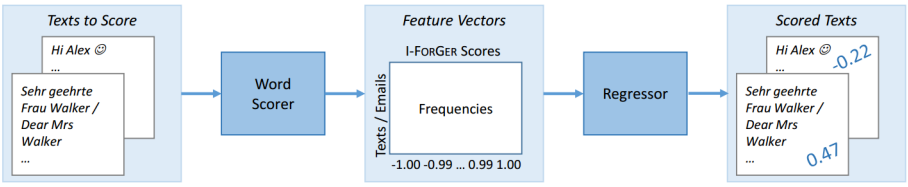

Figure 5: Overview of our workflow to score emails for formality using I-FORGER scores

Spearman's ρ of 0.9198 ($+/-$ 0.0043) over 100 trials. The resulting scores on an informal-formal scale from -1 (informal) to $+1$ (formal) served as basis for our experiments.

5.2 Distribution of I-FORGER Scores

Under the assumption that formal emails include more formal words and informal emails more informal terms, we, first, examined the distribution of scores calculated for the emails' words with the BOOSTED FFNN model learned on I-FORGER. We split our dataset tentatively in two folds: emails with scores from -1 to 0 formed the informal part, whereas emails rated with positive numbers in a range from 0 to $+1$ were regarded as formal. Figure 4 indicates that, in comparison, informal emails indeed contain more negatively scored terms and formal emails comprise more words in the upper part of the informal-formal scale.

5.3 Formality Scoring of Emails

In the final evaluation setup, we tested whether word formality scoring works better than lexicon look-up in traditional resources and whether categorized items or continuous scores get better results. To determine the proper features for a linear regressor predicting formality scores,[21] we used a vector comprising the relative frequencies of an email's word scores as input (count per score divided by the total number of scored words). In one setting, the I-FORGER word scorer tagged (unseen) nouns, finite verbs and adjectives (Figure 5 depicts the workflow for this experiment.). In another setting, we only counted the scores of words already present in I-FORGER (without acquisition step). Besides relative score frequencies, we also tested taking the average score per document (sum of all calculated scores divided by the total number of scored words) as input feature[22] for both settings.

For a comparison of scores against pre-specified categories, we mapped the scores of I-FORGER to formality categories. We divided the scale into five distinct sections (e.g., scores between 0.6 and 1.0 form one category), assigned the respective category to each score and used a classifier instead of a linear regressor to learn the categories of new words. The relative frequencies of the categories then served as input for the linear regressor. We also experimented with ignoring OOV words and only utilizing lexicon look-up for the categorical scenario. For this setting, we exploited the *complete* pre-categorized word lists we got the SWEAR WORDS, COLLOQUIAL WORDS and ELEVATED WORDS from in order to increase coverage. In this way, in case of swear words, e.g., we did not only use the 500 items assembled in the I-FORGER lexicon, but used a list of more than 13,000 entries. As features instead of scores we counted the frequency of swear words, colloquial words and elevated words separately in each email and divided it by the total number of words found in the lexicons.

Table 4 summarizes our results. Scoring words based on I-FORGER yielded significantly better results than any other configuration reaching a strong Spearman's ρ of 0.728. When using the average score per document, there is still a positive correlation with the emails' formality scores. Utilizing a fixed set of lexical terms and not scoring new

Lexicon	OOV	Features	ρ
I-FORGER	Scored	counts	**0.728**
I-FORGER	Scored	average	0.587*
I-FORGER	Ignored	counts	0.446*
I-FORGER	Ignored	average	0.123*
I-FORGER$_{cat}$	Classified	counts	0.476*
CATEGORIES	Ignored	counts	0.335*

Table 4: Spearman's ρ for different configurations (10-fold cross-validation on formality scored CODE ALLTAG$_{S+d}$); significance differences in respect to best model calculated with two-sided Wilcoxon signed-rank test are marked with '*' for $p < 0.005$

[21] We used a neural network with two hidden layers (128 and 64 units) and the same configurations in the KERAS library in TENSORFLOW as reported for NN$_{2Hidden}$ or NN$_{3Hidden}$.

[22] Using the average scores directly to determine a correlation to the emails' formality scores gave comparable results.

words also performed better with score frequencies than using the average. However, compared to employing a word scorer for unseen words, the results for simple lexicon look-up are lower, a finding that seems to be due to the limited coverage of I-FORGER. Therefore, we can conclude that our way of scoring potentially unseen words is an effective and advantageous alternative to using fixed-size, and thus limited, lexical resources.

Employing the relative frequencies of formality categories instead of scores also yielded lower results for both settings, classifying new words (see I-FORGER$_{cat}$) and utilizing lexicon look-up with pre-categorized items (CATEGORIES). This demonstrates the benefit of a scaling approach instead of relying on coarse-grained categories.

6 Conclusion

Different levels of formality these days find increasing attention, both in methodological approaches and NLP applications. The necessity of choosing a socially appropriate tone is particularly evident in digitally mediated discourse, e.g., formal business or informal private email communication (Chhaya et al., 2018) or social media interaction via reviews, chats, or blogs (Pavlick and Tetreault, 2016; Gonzàlez Bermúdez, 2015). The increasing relevance of conversationally adequate virtual personal assistants (Shamekhi et al., 2016), chatbots (Chaves et al., 2019) and automatic procedures for smart response generation (Kannan et al., 2016) requires sensitivity on the generator's side to strike the right tone and avoid the false one. Similarly, machine translation poses special problems when expressions of (in)formality have to be adequately transferred between different languages (Niu et al., 2018). Progress in monitoring formality levels is a methodological prerequisite for several downstream applications that have to comply with users' habitual expectations or increase user satisfaction, e.g., in commercial interactions (customer service communication) (Liebrecht et al., 2020; Elsholz et al., 2019) or medical consultation (Fadhil and Schiavo, 2019).

As a methodological contribution, we here propose a lexical approach to computational style analysis based on I-FORGER, a lexicon whose (3,000) items are scaled on a continuous informal-formal spectrum. We make three new contributions to style analysis: First, a language-independent lexicon acquisition architecture employing sentence embeddings forms the basis for computing sentence similarity, thus finding formality-sensitive lexical items not contained in the seeds. Second, best-worst scaling is used for creating gold standards available for an in-depth intrinsic and extrinsic evaluation of the new lexical resource. Finally, I-FORGER stands out as the first formality-informed lexicon for the German language. This resource is available at https://github.com/ee-2/I-ForGer.

Despite our lexical focus, we are aware of the fact that formality is not only lexically expressed. Consequently, a lexicon-based approach has to be complemented by methods that account for non-lexicalized varieties of formality. Such forms may include syntactic variability, linguistic complexity and readability, as well as correctness of language use regarding orthography, morphology and syntax. For research on formality detection incorporating its syntactic, semantic and discourse facets, cf., e.g., Heylighen and Dewaele (1999), Li et al. (2013) or Pavlick and Tetreault (2016). These branches will also be part of our future work. Still, a (potentially) large portion of formality assessments is rooted in lexical signals, which we capture by the methodology advanced in this paper.

Acknowledgments

We thank the anonymous reviewers of our paper for their remarks, which helped us strengthen the paper. Special thanks go to Michael Wiegand for additional hints and comments.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR : an easy-to-use framework for state-of-the-art NLP. In *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Demonstrations Session. Minneapolis, Minnesota, USA, June 3-4, 2019*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: Main Conference. Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1638–1649.

Salha M. Alzahrani, Naomie Salim, and Ajith Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Transactions on Systems, Man, and Cybernetics — Part C: Applications and Reviews*, 42(2):133–149.

Michael Beißwenger. 2013. Das Dortmunder Chat-Korpus. *Zeitschrift für germanistische Linguistik*, 41(1):161–164.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *HICSS-50 — Proceedings of the 50th Hawaii International Conference on System Sciences 2017. Hawaii, USA, January 4-7, 2017*, pages 2213–2222.

Julian Brooke and Graeme Hirst. 2014. Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *COLING 2014 — Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland, August 23-29, 2014*, pages 2172–2183.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *COLING 2010 — Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China, 23-27 August 2010*, pages 90–98.

Sven Buechel and Udo Hahn. 2018. Word emotion induction for multiple languages as a deep multi-task learning problem. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 1907–1918.

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: identifying appropriate register for chatbot language design. In *HAI '19 — Proceedings of the 7th ACM International Conference on Human-Agent Interaction. Kyoto, Japan, October 6-10, 2019*, pages 102–109.

Niyati Chhaya, Kushal Chawla, Tanya Goyal, Projjal Chanda, and Jaya Singh. 2018. Frustrated, polite or formal: quantifying feelings and tone in emails. In *PEOPLES 2018 — Proceedings of the 2nd Workshop on Computational Modeling of PEople's Opinions, PersonaLity, and Emotions in Social media @ NAACL-HLT 2018. New Orleans, Louisiana, USA, June 6, 2018*, pages 76–86.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: unpaired text style transfer without disentangled latent representation. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, pages 5997–6007.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen W. Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *EACL 2014 — Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, April 26-30, 2014*, pages 702–711.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2019. BERT : pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA, June 2-7, 2019*, volume 1: Long and Short Papers, pages 4171–4186.

Steven H. H. Ding, Benjamin C. M. Fung, Farkhund Iqbal, and William K. Cheung. 2019. Learning stylometric representations for authorship analysis. *IEEE Transactions on Cybernetics*, 49(1):107–121.

Steven Du and Xi Zhang. 2016. AICYBER's system for IALP 2016 Shared Task: character-enhanced word vectors and boosted neural networks. In *IALP 2016 — Proceedings of the [20th] 2016 International Conference on Asian Language Processing. Tainan, Taiwan, November 21-23, 2016*, pages 161–163.

Philipp Dufter and Hinrich Schütze. 2019. Analytical methods for interpretable ultradense word embeddings. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing. Hong Kong, China, November 3-7, 2019*, pages 1185–1191.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2019. At the lower end of language: exploring the vulgar and obscene side of German. In *ALW-3 — Proceedings of the 3rd Workshop on Abusive Language Online @ ACL 2019. Florence, Italy, August, 1, 2019*, pages 119–128.

Elisabeth Eder, Ulrike Krieg-Holz, and Udo Hahn. 2020. CODE ALLTAG 2.0 : a pseudonymized German-language email corpus. In *LREC 2020 — Proceedings of the 12th International Conference on Language Resources and Evaluation. Marseille, France, May 11-16, 2020*, pages 4466–4477.

Maciej Eder, Mike Kestemont, and Jan Rybicki. 2016. Stylometry with R : a package for computational text analysis. *R Journal*, 16(1):107–121.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *ACL 2020 — Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [Seattle, Washington, USA,] July 5-10, 2020 (Virtual Event)*, pages 3154–3160.

Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. 2019. Exploring language style in chatbots to increase perceived product value and user engagement. In *CHIIR '19 — Proceedings of the 2019 Conference on Human Information Interaction and Retrieval. Glasgow, Scotland, UK, March 10-14, 2019*, pages 301–305.

Ahmed Fadhil and Gianluca Schiavo. 2019. Designing for health chatbots. ArXiv preprint arXiv:1902.09022.

Jessica Ficler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *StyVa 2017 — Proceedings of the [1st] Workshop on Stylistic Variation @ EMNLP 2017. Copenhagen, Denmark, September 8, 2017*, pages 94–104.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: exploration and evaluation. In *AAAI-IAAI-EAAI '18 — Proceedings of the 32nd AAAI Conference on Artificial Intelligence & 30th Conference on Innovative Applications of Artificial Intelligence & 8th Symposium on Educational Advances in Artificial Intelligence. New Orleans, Louisiana, USA, February 2-7, 2018*, pages 663–670.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *NAACL-HLT 2019 — Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota, USA, June 2-7, 2019*, pages 3168–3180.

Meritxell Gonzàlez Bermúdez. 2015. An analysis of Twitter corpora and the differences between formal and colloquial tweets. In *TweetMT 2015 — Proceedings of the Tweet Translation Workshop 2015 @ SEPLN 2015. Alicante, Spain, September 15, 2015*, number 1445 in CEUR Workshop Proceedings, pages 1–7.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. 2018. Learning word vectors for 157 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 3483–3487.

Hans van Halteren, R. Harald Baayen, Fiona J. Tweedie, Marco Haverkort, and Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12:65–77.

Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. 2011. Language of vandalism: improving WIKIPEDIA vandalism detection via stylometric analysis. In *ACL-HLT 2011 — Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Portland, Oregon, USA, 19-24 June 2011*, volume 2: Short Papers, pages 83–88.

Benjamin Heinzerling and Michael Strube. 2018. BPEMB: tokenization-free pre-trained subword embeddings in 275 languages. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 2989–2993.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. Technical report, Center "Leo Apostel", Free University of Brussels.

Leo Hickey. 1993. Stylistics, pragmatics and pragmastylistics. *Revue Belge de Philologie et d'Histoire*, 71(3):573–586.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representation learning for non-parallel text style transfer. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, pages 424–434.

Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufman, Andrew D. Tomkins, Balint Miklos, Gregory S. Corrado, László Lukács, Marina Ganea, Peter Young, and Vivek Ramavajjala. 2016. Smart reply: automated response suggestion for email. In *KDD 2016 — Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, August 13-17, 2016*, pages 955–964.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 811–817.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best-worst scaling more reliable than rating scales: a case study on sentiment intensity annotation. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, British Columbia, Canada, July 30 - August 4, 2017*, volume 2: Short Papers, pages 465–470.

Shibamouli Lahiri. 2015. SQUINKY! A corpus of sentence-level formality, informativeness, and implicature. CoRR, abs/1506.02306.

Shibamouli Lahiri, Prasenjit Mitra, and Xiaofei Lu. 2011. Informality judgment at sentence level and experiments with formality score. In *Computational Linguistics and Intelligent Text Processing. CICLing 2011 — Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing. Tokyo, Japan, February 20-26, 2011*, number 6609 in Lecture Notes in Computer Science (LNCS), pages 446–457. Springer.

Guillaume Lample, Sandeep Subramanian, Eric M. Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *ICLR 2019 — Proceedings of the 7th International Conference on Learning Representations. New Orleans, Louisiana, USA, May 6-9, 2019*.

Haiying Li, Zhiqiang Cai, and Arthur C. Graesser. 2013. Comparing two measures for formality. In *FLAIRS 2013 — Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference. St. Pete Beach, Florida, USA, May 22–24, 2013*, pages 220–225.

Juncen Li, Robin Jia, He He, and Percy S. Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, pages 1865–1874.

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui. 2017. Inferring affective meanings of words from word embedding. *IEEE Transactions on Affective Computing*, 8(4):443–456.

Christine Liebrecht, Lena Sander, and Charlotte Van Hooijdonk. 2020. Too informal? How a chatbot's communication style affects brand attitude and quality of interaction. In *Conversations 2020 — Proceedings of the 4th International Workshop on Chatbot Research. [Amsterdam, Netherlands], 23-24 Nov 2020 (Virtual Event)*, page [16pp.].

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Xu Sun, and Zhifang Sui. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *IJCAI '19 — Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, August 10-16, 2019*, pages 5116–5122.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, December 5-10, 2013*, pages 3111–3119.

Alejandro Mosquera and Paloma Moreda. 2012. SMILE : an informality classification tool for helping to assess quality and credibility in Web 2.0 texts. In *RAMSS 2012 — Proceedings of the 1st Workshop on Real-Time Analysis and Mining of Social Streams @ ICWSM 2012. Dublin, Ireland, June 4, 2012*, number WS-12-02 in AAAI Technical Report, pages 2–7.

Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):#86.

Xing Niu, Sudha Rao, and Marine Carpuat. 2018. Multi-task neural models for translating between styles within and across languages. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: Main Conference. Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1008–1021.

Bryan Orme. 2009. Maxdiff analysis: simple counting, individual-level logit, and HB. *Sawtooth Software, Inc.*

Antonio Pascucci, Raffaele Manna, Ciro Caterino, Vincenzo Masucci, and Johanna Monti. 2020a. Is this hotel review truthful or deceptive? A platform for disinformation detection through computational stylometry. In *STOC 2020 — Proceedings of the 1st International Workshop on Social Threats in Online Conversations: Understanding and Management @ LREC 2020. Marseille, France, May 11, 2020*, pages 35–40.

Antonio Pascucci, Raffaele Manna, Vincenzo Masucci, and Johanna Monti. 2020b. The role of computational stylometry in identifying (misogynistic) aggression in English social media texts. In *TRAC-2 2020 — Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying @ LREC 2020. Marseille, France, May 16, 2020*, pages 69–75.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA, May 31 - June 5, 2015*, pages 218–224.

Ellie Pavlick and Joel R. Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Kelly Peterson, Matt Hohensee, and Fei Xia. 2011. Email formality in the workplace: a case study on the Enron corpus. In *LSM 2011 — Proceedings of the Workshop on Language in Social Media @ ACL-HLT 2011. Portland, Oregon, USA, 23 June 2011*, pages 86–95.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, volume 1: Long Papers, pages 231–240.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan R. Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, pages 866–876.

Sudha Rao and Joel R. Tetreault. 2018. Dear Sir or Madam, may I introduce the GYAFC dataset: corpus, benchmarks and metrics for formality style transfer. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 129–140.

Nils Reimers and Iryna Gurevych. 2019. SENTENCE-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing. Hong Kong, China, November 3-7, 2019*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. ArXiv preprint arXiv:2004.09813.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *NAACL-HLT 2016 — Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California, USA, June 12-17, 2016*, pages 767–777.

Cícero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. In *ACL 2018 — Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Victoria, Australia, July 15-20, 2018*, volume 2: Short Papers, pages 189–194.

Dietmar Schabus, Marcin Skowron, and Martin Trapp. 2017. One million posts: a data set of German online discussions. In *SIGIR '17 — Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 1241–1244.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL 2016 — Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany, August 7-12, 2016*, volume 1: Long Papers, pages 1715–1725.

Ameneh Shamekhi, Mary Czerwinski, Gloria Mark, Margeigh Novotny, and Gregory A. Bennett. 2016. An exploratory study toward the preferred conversational style for compatible virtual agents. In *Intelligent Virtual Agents. IVA 2016 — Proceedings of the 16th International Conference on Intelligent Virtual Agents. Los Angeles, California, USA, September 20-23, 2016*, number 10011 in Lecture Notes in Artificial Intelligence (LNAI), pages 40–50. Springer.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 2019. Semi-supervised text style transfer: cross projection in latent space. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing. Hong Kong, China, November 3-7, 2019*, pages 4936–4945.

Fadi Abu Sheikha and Diana Z. Inkpen. 2010. Automatic classification of documents by formality. In *NLPKE 2010 — Proceedings of the 6th IEEE International Conference on Natural Language Processing and Knowledge Engineering. Beijing, China, 21-23 August 2010*, page [5pp.].

Fadi Abu Sheikha and Diana Z. Inkpen. 2011. Generation of formal and informal sentences. In *ENLG 2011 — Proceedings of the 13th European Workshop on Natural Language Generation. Nancy, France, 28-30 September 2011*, pages 187–193.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Advances in Neural Information Processing Systems 30 — NIPS 2017. Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, California, USA, December 3-9, 2017*, pages 6833–6844.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GERMEVAL Task 2, 2019 Shared Task on the Identification of Offensive Language. In *KONVENS 2019 — Proceedings of the 15th Conference on Natural Language Processing. Erlangen-Nürnberg, Germany, October 9-11, 2019*, pages 352–363.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in Dutch social media. In *TA-COS 2016 — Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety @ LREC 2016. Portorož, Slovenia, 23 May 2016*, pages 11–17.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *EMNLP-IJCNLP 2019 — Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing & 9th International Joint Conference on Natural Language Processing. Hong Kong, China, November 3-7, 2019*, pages 3573–3578.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2020. Formality style transfer with shared latent space. In *COLING 2020 — Proceedings of the 28th International Conference on Computational Linguistics. [Barcelona, Spain,] December 8-13, 2020 (Virtual Event)*, pages 2236–2249.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a lexicon of abusive words: a feature-based approach. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 1046–1056.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GERMEVAL 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of the GermEval 2018 Workshop @ KONVENS 2018. Vienna, Austria, September 21, 2018*, pages 1–10.

Chen Wu, Xuancheng Ren, Fuli Luo, and Xu Sun. 2019. A hierarchical reinforced sequence operation method for unsupervised text style transfer. In *ACL 2019 — Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, July 28 - August 2, 2019*, pages 4873–4883.

Zichao Yang, Zhiting Hu, Chris Dyer, Eric P. Xing, and Taylor Berg-Kirkpatrick. 2018. Unsupervised text style transfer using language models as discriminators. In *Advances in Neural Information Processing Systems 31 — NeurIPS 2018. Proceedings of the 32nd Annual Conference on Neural Information Processing Systems. Montréal, Québec, Canada, December 3-8, 2018*, pages 7298–7309.

Xiaoyuan Yi, Zhenghao Liu, Wenhao Li, and Maosong Sun. 2021. Text style transfer via learning style instance supported latent space. In *IJCAI-PRICAI 2020 — Proceedings of the 29th International Joint Conference on Artificial Intelligence & 17th Pacific Rim International Conference on Artificial Intelligence. [planned: Yokohama, Japan, 11-17 July 2020], 7-15 January 2021 (Virtual Event)*, pages 3801–3807.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. In *ACL 2020 – Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [Seattle, Washington, USA,] July 5-10, 2020 (Virtual Event)*, pages 3221–3228.