Knowledge discovery from patient forums: gaining novel medical insights from patient experiences

Dirkson, A.R.

PART I:

PREPROCESSING MEDICAL SOCIAL MEDIA TEXT

Noise! Noise! Noise!
That’s one thing he hated! The NOISE!

Dr. Suess, *How the Grinch stole Christmas!*

2

Data-driven spelling correction

Edited from: **Anne Dirkson**, Suzan Verberne, Abeed Sarker & Wessel Kraaij (2019), *Data-Driven Lexical Normalization for Medical Social Media*, Multimodal Technologies and Interaction 3(3): 60.

The extraction of knowledge from medical social media is complicated by colloquial language use and misspellings. This noisiness can be reduced through lexical normalization: the transformation of non-standard text to a standardized vocabulary. Yet, lexical normalization of such data has not been addressed effectively.

To this end, we present a data-driven lexical normalization pipeline with a novel spelling correction module for medical social media. We find that our method significantly outperforms state-of-the-art spelling correction methods and can detect mistakes with an F_1 of 0.63 despite an extreme imbalance in the data.

Additionally, we present the first corpus for spelling mistake detection and correction in a medical patient forum. We make this corpus publicly available for the community to facilitate further research on this topic.

2.1. INTRODUCTION

In recent years, user-generated data from social media that contains information about health, such as patient forum posts or health-related tweets, has been used extensively for medical text mining and information retrieval (IR) [116]. This user-generated data encapsulates a vast amount of knowledge, which has been used for a range of health-related applications, such as the tracking of public health trends [267] and the detection of adverse drug events [266]. However, the extraction of this knowledge is complicated by non-standard and colloquial language use, typographical errors, phonetic substitutions, and misspellings [65, 229, 261]. This general noisiness of social media text is only aggravated by the complex medical domain [116].

The noisiness of medical social media can be reduced through lexical normalization: the transformation of non-standard text to a standardized vocabulary. Nonetheless, lexical normalization for medical social media has not been explored thoroughly. Medical lexical normalization methods (i.e., abbreviation expansion [210] and spelling correction [168, 230]) have mostly been developed for clinical records or notes. Although clinical records also contain many domain-specific abbreviations and misspellings, their contents are typically focused solely on the medical domain. In contrast, social media text typically covers a wider vocabulary including colloquial language and layman's terms for medical concepts [116, 352]. For medical social media, some recent studies have explored the related task of concept normalization (i.e., the mapping of tokens to standardized concept IDs in an ontology) [116].[1] Community-driven research on the topic has been boosted by the public release of relevant annotated data sets.[2] However, these available annotated data sets for concept normalization do not annotate misspellings explicitly and are thus not suitable for evaluating lexical normalization. As of yet, there are no publicly available annotated data sets for lexical normalization in medical social media.

Currently, the most comprehensive benchmark for lexical normalization in *general-domain* social media is the ACL W-NUT 2015 shared task[3] [19]. The current state-of-the-art system for this task is MoNoise [318]. However, this system is supervised and uses a lookup list of all replacement pairs in the training data as one of its important features. The training data from the task consists of 2,950 tweets with a total of 3,928 annotated non-standard words [19]. As extensive training data is unavailable for medical social media, such supervised systems cannot be employed in this domain. The best unsupervised system available is a modular pipeline with a hybrid approach to spelling, developed by Sarker [261]. Their pipeline also includes a customisable back-end module for domain-specific normalization. However, this back-end module relies on (i) a standard dictionary supplemented manually with domain-specific terms to detect mistakes and (ii) a language model of distributed word representations (word2vec) of generic Twitter data to correct these mistakes (for more detail see Section 2.3.2). For domains that have many out-of-vocabulary (OOV) terms compared to the available dictionaries and language models,

[1] For example, lexical normalization of 'pounding hed' would output 'pounding head', whereas concept normalization would aim to map it to the concept of Headache in a medical ontology such as SNOMED CT. A major difference between lexical and concept normalization is that the latter is constrained to terms of a pre-defined category (e.g., symptoms), whereas lexical normalization is unconstrained and can include any term.

[2] CADEC [151], PsyTAR [353] and the shared tasks of the SMM4H task [268, 335]

[3] `https://noisy-text.github.io/norm-shared-task.html`

such as medical social media, this is problematic.

Manual creation of specialized dictionaries is an unfeasible alternative: medical social media can be devoted to a wide range of different medical conditions and developing dictionaries for each condition (including laymen terms) would be very labor-intensive. Additionally, there are many different ways of expressing the same information and the language use in the forum evolves over time. In this chapter, we present an alternative: a corpus-driven spelling correction approach. Our method is designed to be conservative and to focus on precision to mitigate one of the major challenges of correcting errors in domain-specific data: the loss of information due to the erroneous correction of already correct domain-specific terms. Although dictionary-based retrieval will capture all mistakes, because any word that is not in the dictionary is considered a mistake, thereby attaining a high recall, its precision will be low. This is a result of words that are correct but not present in the dictionary as they will be erroneously marked as mistakes. Many domain-specific terms will fall in this category. In contrast, data-driven methods can capture patterns to recognize these non-mistakes as correct words and thereby improve precision, while recall could go down as these patterns might miss mistakes, for example because they are common. A data-driven detection approach will thus be more precise than dictionary-based retrieval.

In this chapter, we address two research questions:

1. To what extent can corpus-driven spelling correction reduce the out-of-vocabulary rate in medical social media text?

2. To what extent can corpus-driven spelling correction improve the accuracy of health-related classification tasks with social media text?

Our contributions are (1) an unsupervised data-driven spelling correction method that works well on specialized domains with many OOV terms without the need for a specialized dictionary[4] and (2) the first corpus for evaluating mistake detection and correction in a medical patient forum.[5]

The rest of the paper is organized as follows: In Section 2.2, we briefly review related work. In Section 2.3, we discuss the data sets we employ (Section 2.3.1) followed by a detailed description of our methodology (Section 2.3.2). In Section 2.4, we present our evaluation results, which are discussed further in Section 2.5. Lastly, in Section 2.6 we conclude our paper with final insights and an outline of future work.

2.2. Related work

2.2.1. Challenges in correcting spelling errors in medical social media

A major challenge for correcting spelling errors in small and highly specialized domains is a lack of domain-specific resources. This complicates the automatic creation of relevant dictionaries and language models. Moreover, if the dictionaries or language models are not domain-specific enough, there is a high probability that specialized terms will be

[4] Our lexical normalization pipeline is available at: `https://github.com/AnneDirkson/LexNorm`

[5] The corpus is available at `https://github.com/AnneDirkson/SpellingCorpus`

incorrectly marked as mistakes. Consequently, essential information may be lost as these terms are often key to knowledge extraction tasks (e.g., a drug name) and to specialized classification tasks (e.g., does the post contain a side effect of drug X?).

This challenge is further complicated by the dynamic nature of language on medical social media: in both the medical domain and social media novel terms (e.g., novel drug names) and neologisms (e.g., group-specific slang) are constantly introduced. Unfortunately, professional clinical lexicons are also unsuited for capturing the domain-specific terminology on forums, because laypersons and health care professionals express health-related concepts differently [348]. Another complication is the frequent misspellings of key medical terms, as medical terms are typically difficult to spell [352]. This results in an abundance of common mistakes in key terms, and thus, a large amount of lost information if these terms are not handled correctly.

2.2.2. LEXICAL NORMALIZATION OF SOCIAL MEDIA

The emergence of social networks and text messaging has redefined spelling correction to the broader task of lexical normalization, which may also involve tasks like abbreviation expansion [19]. In earlier research, text normalization for social media was mostly unsupervised or semi-supervised (e.g., [121]) due to a lack of annotated data. These methods often pre-selected and ranked correction candidates based on phonetic or lexical string similarity [120, 121]. Han and Baldwin [120] additionally used a trigram language model trained on a large Twitter corpus to improve correction. Although these methods did not rely on training data to correct mistakes, they did rely on dictionaries to determine whether a word *needed* to be corrected [120, 121]. The opposite is true for modern supervised methods: they do not rely on dictionaries but do rely on training data for both misspelling detection and correction. For instance, the best performing method at the ACL W-NUT shared task of 2015 used canonical forms in the training data to develop their own normalization dictionary [144]. Other competitive systems were also supervised and often used deep learning to detect and correct mistakes [175, 208] (for more detail on W-NUT systems see Baldwin et al. [19]). More recent competitive results for this shared task include MoNoise [318]. As mentioned, this system is also supervised and uses a lookup list of all replacement pairs in the training data as an important feature in their spelling correction. Since such specialized resources (appropriate dictionaries or training data) are not available for medical forum data, a method that relies on neither is necessary. We address this gap in this chapter.

Additionally, recent approaches (e.g., [261]) often make use of language models for spelling correction. Language models, however, require a large corpus of comparable text from the same genre and domain [261], which is a major obstacle for employing such an approach in niche domains. Since forums are often highly specialized, the resources that could capture the same language use are limited. Nevertheless, if comparable corpora are available, language models can contribute to effectively reducing spelling errors in social media [261] due to their ability to capture the context of words and to handle the dynamic nature of language.

Recent developments in the NLP field towards distributional language models based on byte-pair (BPE) or character-level encoding instead of word-level encoding call into question the need for prior spelling correction. In general, character-level models are more robust to noise than word-level models, as they can exploit the remaining character structure regardless of errors. Niu et al. [218] recently developed a character-level attentional network model for medical concept normalization in social media which can alleviate the problem of out-of-vocabulary (OOV) terms by using a character-level encoding. Their model is robust to misspellings resulting from double characters, swapping of letters, adding hashtags and deletions.

However, firstly, the robustness to noise of character-based models appears to rely on whether they have been trained on noisy data [26, 132]. Otherwise, they are prone to breaking when presented with synthetic or natural noise [26, 132]. Thus, if sufficiently large amounts of data with similar types of noise are available, character-based models may negate the need for spelling correction. However, in domains lacking such resources, spelling correction in the pre-processing stage is still needed. Secondly, character-based models have computational disadvantages: their computational complexity is higher and it becomes harder to model long-range dependencies [132]. Alternatively, word embeddings designed to be robust to noise [196] could be used. Yet, also for this method, sufficiently large amounts of comparable noisy data are necessary. To provide an indication, Malykh et al. [196] use the Reuters CV-1 corpus consisting of 800,000 news stories ranging from a few hundred to several thousand words in length [177] to generate their robust English word embeddings.

2.2.3. Lexical normalization of clinical records

Like medical social media, clinical notes made by doctors are user-generated and noisy. In fact, Ruch et al. [255] reported about one spelling error per five sentences. Yet, most normalization research for clinical notes has focused on concept normalization instead of lexical normalization [116]. A prominent shared task for concept normalization of clinical notes is Task 2 of the CLEF e-Health workshop in 2014. Its aim was to expand abbreviations in clinical notes by mapping them to the UMLS database [210]. The best system by Wu et al. [343] applied four different trained tagging methods depending on the frequency and ambiguity of abbreviations. Unfortunately, the abbreviations used by doctors are not the same as the ones used by patients, and thus these methods do not transfer.

To correct misspellings in clinical notes, Lai et al. [168] developed a spell checker based on the noisy channel model by Shannon [273]. Noisy channel models interpret spelling errors as distortions of a signal by noise. The most probable message can then be calculated from the source signal and noise models. This is how spelling correction is modeled traditionally [64]. Although their correction accuracy was high, their method relied on an extensive dictionary compiled from multiple sources to detect mistakes. Similarly, the method by Patrick et al. [230] also used a compiled dictionary for detecting errors. For correction, Patrick et al. [230] used edit distance-based rules to generate suggestions which were ranked using a trigram model. Fivez et al. [110] was the first to leverage contextual information to correct errors in clinical records. They developed an unsupervised, context-sensitive method that used word and character embeddings to correct spelling errors. Their approach outperformed the method proposed by Lai et al. [168] for the benchmark MIMIC-III [146]. However, they did not perform any mistake detection, as they simply tried to correct the annotated misspellings of MIMIC-III. In conclusion, the methods developed for spelling correction in clinical records either only focus on correction or rely solely on extensive, compiled dictionaries to find mistakes. Therefore, they are not applicable in domains lacking such resources.

2.3. Materials and methods

2.3.1. Data

Data collection For evaluating spelling correction methods, we use an international patient forum for patients with Gastrointestinal Stromal Tumor (GIST). It is moderated by GIST Support International (GSI). This data set was donated to Dr. Verberne by GSI in 2015. We use a second cancer-related forum to assess the generalisability of our methods: a sub-reddit community on cancer, dating from 16/09/2009 until 02/07/2018.[6] It was scraped using the Pushshift Reddit API.[7] The data was collected by looping over the timestamps in the data. This second forum is roughly four times larger than the first in terms of the number of tokens (See Table 2.1).

Table 2.1: Raw data without punctuation. IQR: Inter-quartile range

	GIST forum	Reddit forum
# Tokens	1,255,741	4,520,074
# Posts	36,277	274,532
Median post length (IQR)	20 (35)	11 (18)

Data annotation Spelling mistakes were annotated for 1000 randomly selected posts from the GIST data. Each token was classified as a mistake (1) or not (0) by the first author. For the first 500 posts, a second annotator checked if any of the mistakes were false positives. In total, 99 of the 109 non-word spelling errors were annotated for correction experiments. The remaining 10 errors were found later during error detection experiments and were therefore only included in these experiments. The corrections for the 53 unique mistakes present in the first 500 posts were annotated individually by two annotators, of which one was a GIST patient and a forum user. Annotators were provided with the complete post to determine the correct word. The initial absolute agreement was 89.0%. If a consensus could not be reached, a third assessor was used to resolve the matter. The remaining mistakes were annotated by the first author. For the correction 'reoccurrence', the synonym 'recurrence' was also considered correct. As far as we are aware, no other spelling error corpora for this domain are publicly available.

To tune the similarity threshold for the optimal detection of spelling mistakes, we used 60% of the annotated data as a development set. The split was done per post and stratified on whether a post contained mistakes or not. Since the data is extremely unbalanced, we balanced the training data to some extent by combining the mistakes with a ten-fold of random correct words with the same word length distribution (see Table 2.2). These words were not allowed to be numbers, punctuation, or proper nouns, because these are ignored by our error detection process. The development set was split in a stratified manner into 10 folds for cross-validation.

[6] `http://www.reddit.com/r/cancer`

[7] `https://github.com/pushshift/api`

Table 2.2: Annotated data for spelling detection experiments. *excluding punctuation, numbers and proper nouns.

	Mistakes (%)	Total word count*
Training set	57 (9.1%)	627
Test set	45 (0.42%)	10760

Corpus for calculating weighted edit matrix Since by default all edits are weighted equally when calculating Levenshtein distance, we needed to compute a weighted edit matrix in order to assign lower costs and thereby higher probabilities to edits that occur more frequently in the real world. We based our weighted edit matrix on a corpus of frequencies for 1-edit spelling errors compiled by Peter Norvig.[8] This corpus is compiled from four sources: (1) a list of misspellings made by Wikipedia editors, (2) the Birkbeck spelling corpus, (3) the Holbrook corpus and (4) the Aspell error corpus.

Specialized vocabulary for OOV estimation in cancer forums To be able to calculate the number of out-of-vocabulary terms in the two cancer forums, a specialized vocabulary was created by merging the standard English lexicon CELEX [46] (73,452 tokens), the NCI Dictionary of Cancer Terms [215] (6,038 tokens), the generic and commercial drug names from the RxNorm [314] (3,837 tokens), the ADR lexicon used by Nikfarjam et al. [217] (30,846 tokens) and our in-house domain-specific abbreviation expansions (DSAE) (42 tokens) (see 2.3.2 for more detail). As many terms overlapped with those in CELEX, the total vocabulary consisted of 118,052 tokens (62.2% CELEX, 5.1% NCI, 26.1% ADR, 6.5% RxNorm and <0.01% DSAE).

2.3.2. METHODS

Preprocessing URLs and email addresses were replaced by the strings -URL- and -EMAIL- using regular expressions. Furthermore, text was lower-cased and tokenized using NLTK. The first modules of the normalization pipeline of Sarker [261] were employed: converting British to American English and normalizing generic abbreviations (see Figure 2.1). Some forum-specific additions were made: Gleevec (British variant: Glivec) was included in the British-American spelling conversion, one generic abbreviation expansion that clashed with a domain-specific one was substituted (i.e., 'temp' defined as *temperature* instead of *temporary*), and two problematic medical terms were removed from the slang dictionary (i.e., 'ill' corrected to 'i'll' and 'chronic' corrected to 'marijuana').

Moreover, the abbreviations dictionary by Sarker [261] was lower-cased. As apostrophes in contractions are frequently omitted in social media posts (e.g., im instead of i'm), we expanded contractions to their full form (e.g., i am). Firstly, contractions with apostrophes were expanded and subsequently those without apostrophes were expanded only if they were not real words according to the CELEX dictionary. Lastly, domain-specific abbreviations were expanded with a lexicon of domain-specific abbreviation expansions (DSAE). The abbreviations were manually extracted from 500 randomly selected posts of the GIST forum data. This resulted in 47 unique abbreviations. Two annotators, of which

[8] http://norvig.com/ngrams/count_1edit.txt

one was a domain expert, individually determined the correct expansion term for each abbreviation, with an absolute agreement of 85.4%. Hereafter, they agreed on the correct form together.[9]

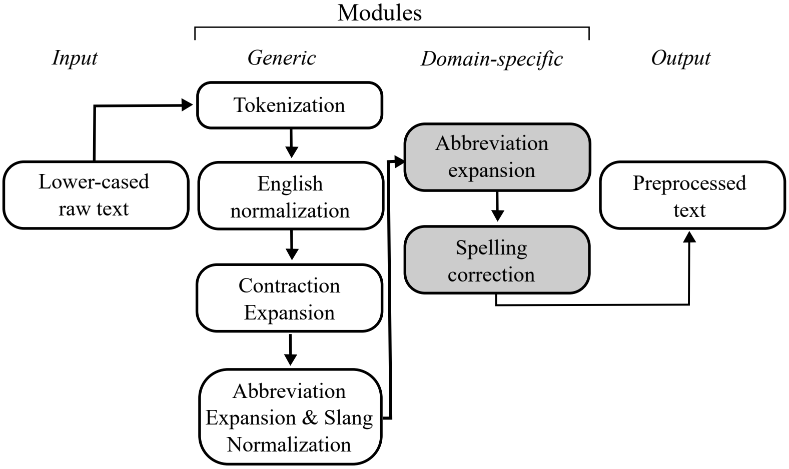


Figure 2.1: Sequential processing pipeline

Spelling correction

Baseline methods We used the method by Sarker [261] as a baseline for spelling correction. Their method combines normalized absolute Levenshtein distance with Metaphone phonetic similarity and language model similarity. For the latter, distributed word representations (skip-gram word2vec) of three large Twitter data sets were used. In this chapter, we used the largest available version of the DIEGO LAB Drug Chatter Corpus (around 1 Billion tweets) [263], as it was the only health-related corpus of the three. We also use a purely data-driven spelling correction method for comparison: Text-Induced Spelling Correction (TISC) developed by Reynaert [248]. It compares the anagrams of a token to those in a large corpus of text to correct mistakes. These two methods are compared with simple absolute and relative Levenshtein distance and weighted versions of both. To evaluate the spelling correction methods, the accuracy (i.e., the percentage of correct corrections) was used. The weights of the edits for weighted Levenshtein distance were computed using the log of the frequencies of the Norvig corpus. We used the log to ensure that a 10x more frequent error does not become 10x as cheap, as this would make infrequent errors too improbable. In order to make the weights inversely proportional to the frequencies and scale the weights between 0 and 1 with lower weights signifying lower

[9] This abbreviations lexicon is shared at `https://github.com/AnneDirkson/LexNorm`

costs for an edit, the following transformation of the log frequencies was used: Weight Edit Distance = $\frac{1}{1+log(frequency)}$.

Correction candidates Spelling correction methods were first compared using the terms from the specialized vocabulary for cancer forums (see section 2.3.1) as correction candidates. This enables us to evaluate the methods independently of the vocabulary present in the data. Hereafter, we assessed the impact of using correction candidates from the data itself, since our aim is to develop a method that is independent of manually compiled lexicons. Numbers, proper nouns, and punctuation are ignored as possible correction candidates.

We inspected whether restricting the pool of eligible correction candidates based on their corpus frequency relative to that of the token aids correction. We use relative corpus frequency thresholds ranging from at least 0 times (no restriction) to 10 times more frequent than the token. The underlying idea is that the correct word will be used more often than the incorrect word and by restricting the candidates we prevent implausible but similar words from hindering correction. This, for instance, prevents mistakes from being corrected by other similar and roughly equally frequent mistakes. A relative, instead of absolute, threshold that depends on the frequency of the mistake enables us to also correct mistakes even if they occur more commonly (e.g., misspellings of a complex medication name). Candidates are considered in order of frequency. Of the candidates with the highest similarity score, the first is selected.

We tried two different approaches to further improve correction by altering the pool of correction candidates. Firstly, we tested whether prior lemmatization of the spelling errors with or without prior lemmatization of the correction candidates could improve spelling correction. Secondly, we investigated the effect of imposing an additional syntactic restriction on the correction candidates, namely only allowing those with the same Part-of-Speech tag at least once in the data or the same surrounding POS tags to the left and right (i.e., the POS context) at least once in the data. McNemar tests were used to test whether the predictions of various correction methods are significantly different. In all follow-up experiments, correction candidates were derived from the respective data set and constrained by the optimal relative corpus frequency threshold.

Improving the baseline method For the best baseline method with data-driven candidates, we explored whether the context of the token could aid the correction further using (1) language models of the forum itself or (2) publicly available distributed and sequential language models of health-related social media data. This last category includes the distributed word2vec (dim= 400) and sequential trigram language models developed by Sarker and Gonzalez [263] and the distributed word2vec (dim = 200) HealthVec model developed by Miftahutdinov et al. [207]. The models by Sarker and Gonzalez [263] are based on around 1 billion Twitter posts derived from user timelines where at least 1 medication is mentioned. A smaller version of this language model is used in the current state-of-the-art normalization pipeline for general social media [261].[10] The HealthVec model is based on the Health Dataset consisting of around 2.5 million

[10] Language models can be obtained from: `https://data.mendeley.com/datasets/dwr4xn8kcv/3`

user comments from six web resources: WebMD, Askapatient, patient.info, Dailystrength, drugs.com, and product reviews from the Amazon Dataset.[11] Besides employing these language models, we explored whether adding double Metaphone phonetic similarity [233] improves correction. Phonetic similarity is a measure of how phonetically similar an error is to the potential correction candidate.

The best baseline method was combined with these similarity measures (i.e., phonetic similarity or the similarity based on a language model) in a weighted manner with weights ranging from 0 to 1 with steps of 0.1. The inverse weight was assigned to the baseline similarity measure. For all language models, if the word was not in the vocabulary, then the model similarity was set to 0, essentially rendering the language model irrelevant in these cases. To investigate the impact of adding these contextual measures, Pearson's r is used to calculate the correlation between the correction accuracy and the assigned weight.

Correcting Concatenation Errors If a word is not in the Aspell dictionary[12], but is also not a spelling mistake, our method checks if it needs to be split into two words. It is split only if it can be split into two words of at least 3 letters which both occur more in the corpus more frequently than the relative corpus frequency boundary. For each possible split, the frequency of the least frequent word is considered. The most plausible split is the one for which this lower frequency is the highest (i.e., the least frequent word occurs the most). Words containing numbers (e.g., 3months) are the exception: they are split so that the number forms a separate word.

Spelling mistake detection We manually constructed a decision process, inspired by the work by Beeksma et al. [25], for detecting spelling mistakes (See Figure 2.7). The optimal relative corpus frequency threshold determined for spelling correction in our earlier experiments is adopted. On top of this threshold, the decision process uses the similarity of the best candidate to the token to identify mistakes. If there is no similar enough correction candidate available, then the word is more likely to be a unique domain-specific term we do not wish to correct than a mistake. The minimum similarity threshold is optimized with a 10-fold cross validation grid search from 0.40 to 0.80 (steps of 0.02). The loss function used to tune the parameters was the $F_{0.5}$ score, which places more weight on precision than the F_1 score. We believe it is more important to not alter correct terms than to retrieve incorrect ones. Candidates are considered in order of frequency. Of the candidates with the highest similarity score, the first is selected. The error detection automatically labels numbers, punctuation, proper nouns, and words present in the Aspell dictionary as correct. We used the word list 60 version of the Aspell dictionary, as is recommended for spelling correction. To verify that medication names were not being classified as proper nouns and thereby excluded from spelling correction, we checked the part-of-speech tags of the most common medication for GIST patients (gleevec) and two of its common misspellings (gleevic and gleevac). For gleevec, 81.4% of the mentions were classified as nouns (NN). The next two largest categories were adjectives (JJ) (7.2%), plural nouns (NNS) (4.7%) and verbs (VB) (3.9%). The remaining 2.8% were divided over 10 POS-tags (ranging from 0.6% to 0.0005%). Most importantly, none were classified as

[11] Available at: `http://jmcauley.ucsd.edu/data/amazon`

[12] Available at: http://Aspell.net/

proper nouns (NNP or NNPS). Similarly, gleevic and gleevac were labeled as nouns (NN) 78.1% and 83.9% of the time and neither was ever labelled as a proper noun. For gleevic, the remaining cases were divided amongst plural nouns (11.4%), adjectives (8.3%) and verbs (2.2%). For gleevac, the remainder was divided between verbs (11.9%) and adjectives (4.2%).

We compared our optimized decision process with and without concatenation error detection (see Section 2.3.2) with error detection using two commonly used dictionaries, CELEX [46] and Aspell, with Microsoft Word and with TISC, another data-driven detection method [248]. Significance was calculated with McNemar tests. Any mistakes overlapping between the training and test set were not included in the evaluation.

Impact of the corpus size on detection To measure the influence of the size of the corpus on spelling mistake detection, we varied the size of the corpus from which correction candidates are derived. The token frequencies of errors and candidates were both calculated using this corpus. Therefore, the frequencies of mistakes and potential corrections would vary and we could estimate for each corpus size how much the error detection in 1000 posts would change. We used Jaccard similarity to measure the overlap between the error predictions of each possible combination of two different corpus sizes.

As our relative corpus frequency threshold is a minimal threshold, bigger corpora and thus larger differences between the token frequency of the error and that of the correct variant would not pose a problem. Consequently, we randomly selected posts to artificially downsize our two cancer forums exponentially. We used sizes ranging from 1000 posts to all forum posts. The 1000 posts for which errors were detected were always included in the corpus. For the GIST forum, we used the 1000 annotated posts.

Impact of the degree of noisiness of the data To investigate the impact of the level of noise in the data on spelling correction and detection, we simulated data sets with varying proportions of misspellings. As our method was designed on data with few errors (< 1% in our sample), this will help us to understand to what extent our method can generalize to more noisy user-generated data. We generated artificial data by altering the number of misspellings in two cancer-related fora.

In line with the work by Niu et al. [218], we generated artificial noise typical of social media text by (i) deleting a single letter, (ii) doubling a letter and (iii) swapping two adjacent letters. Niu et al. [218] also added hashtags to words, but as this is only relevant for Twitter we omit this transformation. Words are randomly selected based on a pre-determined probability of occurrence (1,2,3,4,8 and 16%). Which letter is removed or swapped in the word is dependent on the normalized likelihood of a deletion or swap occurring in real-word data. We use the normalized log frequencies of the Norvig corpus [219]. Additionally, the log frequencies were normalized per word to sum to 1. Which letter is doubled is randomly selected, as frequencies for such operations are not available. We evaluated the spelling correction and detection for each forum with the average of three runs of 1000 randomly selected posts with 3 different seeds.

Effect on OOV rate The percentage of out-of-vocabulary (OOV) terms is used as an estimation of the quality of the data: less OOV-terms and thus more in-vocabulary (IV) terms are a proxy for cleaner data. As the correction candidates are derived from the data itself, one must note that words that are not part of Aspell may also be transformed from IV to OOV. OOV analysis was done manually.

External validation To evaluate the impact of lexical normalization as a preprocessing step on the performance of separate downstream tasks, we perform extrinsic evaluation of our pipeline by running six text classification experiments. We obtained six publicly available health-related Twitter data sets ranging in size from 588 to 16,141 posts (see Table 2.3). As can be seen in Table 2.3, the data sets also have varying degrees of imbalance. It is not uncommon for social media data sets to be highly imbalanced and thus we investigate whether the impact of spelling correction is influenced by imbalance. The data sets were retrieved from the data repository of Dredze[13] and the shared tasks of Social Media Mining for Health Applications (SMM4H) workshop 2019.[14]

Text classification was performed before and after normalization using default sklearn classifiers: Stochastic Gradient Descent (SGD), Multinomial Naive Bayes (MNB) and Linear Support Vector Machines (SVC). Unigrams were used as features. A 10-fold cross-validation was used to determine the quality of the classifiers and a paired t-test was applied to determine significance of the absolute difference. Only the best performing classifier is reported per data set. For the shared tasks of the SMM4H workshop, only the training data was used.

Table 2.3: Six classification data sets of health-related Twitter data. *SMM4H: Social Media Mining for Health Applications workshop

Data set	Task	Size	Positive Class
Task 1 SMM4H 2019*	Presence adverse drug reaction	16,141	8.7%
Task 4 SMM4H 2019* Flu vaccine	Personal health mention of flu vaccination	6,738	28.3%
Flu Vaccination Tweets [141]	Relevance to flu vaccination	3,798	26.4%
Twitter Health [231]	Relevance to health	2,598	40.1%
Task4 SMM4H 2019* Flu infection	Personal health mention of having flu	1,034	54.4%
Zika Conspiracy Tweets [91]	Contains pseudo-scientific information	588	25.9%

To evaluate our method on generic social media text, we used the test set of the ACL W-NUT 2015 task [19]. The test set consists of 1967 tweets with 2024 one-to-one, 704 one-to-many, and 10 many-to-one mappings. We did not need to use the training data, as our method is unsupervised. We omitted the expansion of contractions from our normalization pipeline for the W-NUT task, because expanding contractions was not part of the goals of the task. Error analysis was done manually on the 100 most frequent errors.

[13] http://www.cs.jhu.edu/~mdredze/data/

[14] https://healthlanguageprocessing.org/smm4h/challenge/

2.4. RESULTS

In this section, we will report the distribution of spelling errors in our corpus (2.4.1), the evaluation of spelling correction (2.4.2) and detection methods (2.4.3) on our spelling corpus and the impact of corpus size (2.4.4) and the level of noise in the corpus (2.4.5) on the efficacy of our method. Hereafter, we assess the impact of our method on the OOV rate in two cancer-related fora (2.4.6) and on classification accuracy of six health-related Twitter benchmarks (2.4.7). We also evaluate the performance of our method on the W-NUT shared task for generic social media normalization (2.4.7).

2.4.1. ERROR DISTRIBUTION

Spelling errors can be divided into non-word errors (i.e., errors that are not valid words) and real-word errors (i.e., errors that result in another valid word) [164]. Incorrect concatenations and splits can be either. For example, 'scan' to 'scant' is a real word error whereas 'side effects' to 'sideeffects' is a non-word error. We focus on correcting non-word errors, as we are not interesting in correcting syntactic or semantic errors [164].

Nonetheless, we investigate the prevalence of these error types in the data to gain insight into which types of errors are made in medical social media text. As can be seen in Table 2.4, our corpus of 1000 medical posts from the GIST forum mainly contains non-word errors. Moreover, non-word errors contain the highest percentage of medical misspellings (47.7%). Comparatively, only 20% of real word errors are medical terms. Most posts do not contain any errors (see Figure 2.2), but for those that do, there was in most cases only one error per post.

Table 2.4: Error distribution in 1000 GIST posts

Error type	Non-word	Incorrect splits	Incorrect concatenations	Real word
Amount	109	17	24	30
Non-Medical/Medical	57/52	25/5	14/3	18/6
Percentage of tokens	0.32%	0.05%	0.07%	0.09%
Example mistake	gleevac	gall bladder	sideeffects	scant
Example correction	gleevec	gallbladder	side effects	scan

2.4.2. SPELLING CORRECTION

The normalization step prior to spelling correction (see Figure 2.1) corrected 12 of the 99 spelling errors, such as 'feelin' to 'feeling'. These errors are all on the fuzzy boundary between spelling errors and slang. Thus, spelling correction experiments were performed with the remaining annotated 87 spelling errors.

The state-of-the-art method for generic social media by Sarker [261] performs poorly for medical social media: it corrects only 19.3% of the mistakes (see Table 2.5). In fact, it performed significantly worse ($p < 0.0001$) than all edit distance based methods. Computationally, it is also much slower (see Table 2.6). A second established data-driven approach, TISC [248], performed even more poorly (14.8%). TISC was also significantly worse than all edit-based methods ($p < 0.0001$). Relative weighted edit distance performed the best with an accuracy of 68.2%. The theoretical upper bound

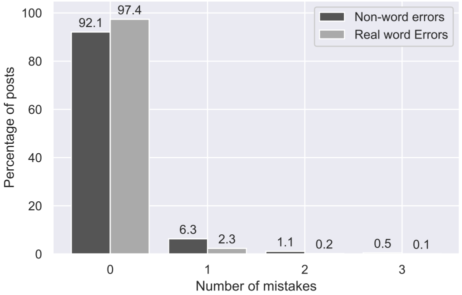


Figure 2.2: Distribution of non-word and real word errors across posts in the GIST forum.

for accuracy was 92.0%, because not all corrections occur in the specialized dictionary. Examples of corrections can be seen in Table 2.7.

Table 2.5: Correction accuracy using a specialized vocabulary. AE: absolute edit distance. RE: relative edit distance. WAE: weighted absolute edit distance. WRE: weighted relative edit distance. *Only the best corpus frequency threshold is reported

Source of candidates	Ceiling	**AE**	**RE**	**WAE**	**WRE**	**Sarker**	**TISC**
Specialized vocabulary	92.0%	58.0%	64.7%	63.3%	68.2%	19.3%	14.8%
GIST forum text*	97.6%	**73.9%**	**73.9%**	70.4%	72.7%	44.3%	-

Table 2.6: Mean computation time over 5 runs

AE	**RE**	**WAE**	**WRE**	**Sarker**
13.36 ms	14.04 ms	29.45 ms	32.00 ms	904.33 ms

However, when using candidates derived from the data itself, unweighted absolute and relative edit distance perform the best. Relative edit distance accurately corrects 73.9% of all mistakes at a relative corpus frequency threshold (θ) of 9, while absolute edit distance does so at a θ of 2 to 5 (See Table 2.5 and Figure 2.3). A θ of 9 means that candidates are only considered plausible if they occur 9 times more frequently than the spelling error. We elect to use relative edit distance, because it is more fine-grained than absolute edit distance, especially for short words. Using data-driven candidates increases the theoretical upper bound from 90.2% to 97.6%. This showcases the limitations of using dictionaries for correction.

Nonetheless, simply using all words from the data as possible candidates (i.e., a corpus frequency threshold of 0) for every spelling error results in a very low correction accuracy (see Figure 2.3). However, imposing the restriction that the corpus frequency of a viable correction candidate must be at least double (2x) that of the mistake, significantly improves correction ($p < 0.0001$) for all correction methods. In that case, for a mistake occurring 10 times, only words occurring at least 20 times are considered. Thus, the

Table 2.7: Corrections by different methods with candidates from a specialized vocabulary. *Gleevec and Sutent are important medications for GIST patients.

Mistake	Correction	AE	RE	WAE	WRE	Sarker	TISC
gleevac	gleevec*	**gleevec**	**gleevec**	**gleevec**	**gleevec**	colonic	gleevac
stomack	stomach	**stomach**	**stomach**	smack	**stomach**	smack	smack
ovari	ovary	**ovary**	**ovary**	**ovary**	**ovary**	ova	atari
sutant	sutent*	mutant	mutant	**sutent**	**sutent**	mutant	dunant
mestastis	metastasis	miscasts	**metastasis**	**metastasis**	**metastasis**	miscasts	mestastis

assumption that corrections are more common than mistakes appears to hold true. However, at any threshold all edit distance based methods still significantly ($p < 0.001$) outperform the state-of-the-art method [261], in line with previous results (Table 2.5). Examples of corrections with data-driven candidates are reported in Table 2.8.

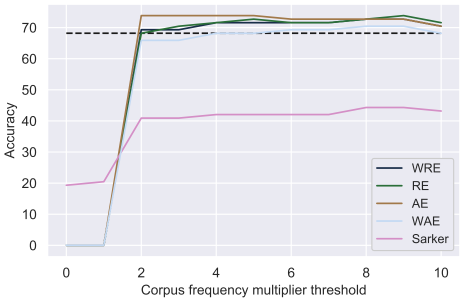


Figure 2.3: Correction accuracy of unique mistakes using correction candidates from the data at various minimum relative corpus frequency thresholds. Dotted line indicates the best correction accuracy using dictionary-derived candidates.

Table 2.8: Corrections by different methods with data-driven candidates. AE: absolute edit distance. RE: relative edit distance. WAE: weighted absolute edit distance. WRE: weighted relative edit distance.

Mistake	Correction	AE	RE	WAE	WRE	Sarker
gleevac	gleevec	**gleevec**	**gleevec**	**gleevec**	**gleevec**	**gleevec**
stomack	stomach	**stomach**	**stomach**	**stomach**	**stomach**	stuck
ovari	ovary	**ovary**	**ovary**	**ovary**	**ovary**	ovarian
sutant	sutent	**sutent**	**sutent**	**sutent**	**sutent**	mutant
mestastis	metastasis	metastis	metastis	metastis	metastis	metastis

The accuracy of the best baseline method, namely relative edit distance with a θ of 9, is unaffected by prior lemmatization of the spelling errors (see Table 2.9). It thus appears that if prior lemmatization can correct the error, our method automatically does so. In contrast, additional lemmatization of their corrections and of the correction candidates significantly reduces accuracy ($p = 0.021$ and $p = 0.011$) compared to omitting prior lemmatization. Thus, lemmatization of the data or candidates prior to spelling correction is not recommended.

NoLemmatization	LemmatizedInput	+ LemmatizedOutput	+ LemmatizedCandidates
73.6%	73.6%	64.7%	67.0%

Table 2.9: Effect of lemmatization of the errors (LemmatizedInput), their corrections (LemmatizedOutput) and correction candidates (LemmatizedCandidates) on spelling correction accuracy using RE (θ= 9)

Adding weighted phonetic similarity Previous research has shown that when users are faced with the task of writing an unfamiliar, complex word like a drug name, they tend to revert to phonetic spelling [235]. Therefore, we investigate whether adding a weighted phonetic component may improve correction. This is not the case: The weight assigned to phonetic similarity has a strong negative correlation (-0.92) with the correction accuracy ($p < 0.0001$) (see Figure 2.4). This suggests that such phonetic errors are already captured by our frequency-based method.

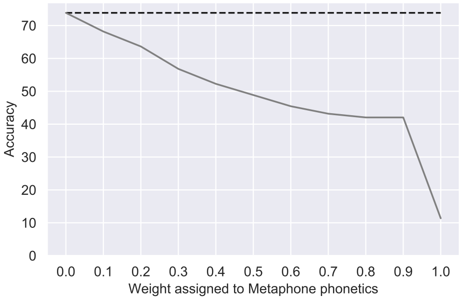
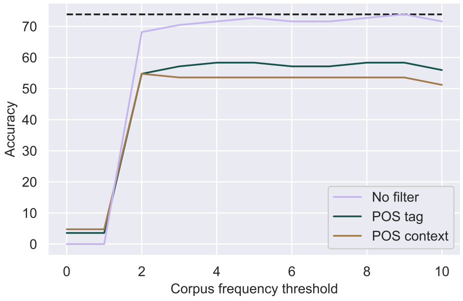


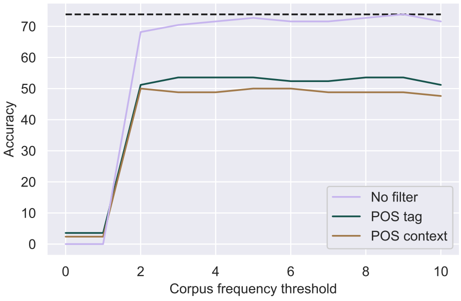
Figure 2.4: Correction accuracy with additional weighted double Metaphone phonetic similarity. Dotted line indicates the best accuracy with relative edit distance alone.

Adding weighted contextual similarity Previous work has indicated that the context of spelling mistakes might be helpful to improve spelling correction [110]. Since domain-specific resources are scarce, one potential approach is to use the contextual information present in the corpus itself. Based on work by Beeksma et al. [25], we tried to use the Part-of-Speech (POS) tag of the error or the POS tags of its neighbors to constraint correction candidates. However, as can be seen in Figure 2.5, adding these constraints reduces correction accuracy, although not significantly. Aside from some additional errors, using POS context as a constraint results in identical errors as enforcing a similar POS tag for potential correction candidates, regardless of whether NLTK or Spacy is used.

As many modern methods use language models to aid spelling correction [261], we also examine whether we can leverage contextual information by using language models of the corpus itself to improve correction accuracy. For both Word2vec and FastText



(a) NLTK POS tags



(b) Spacy POS tags

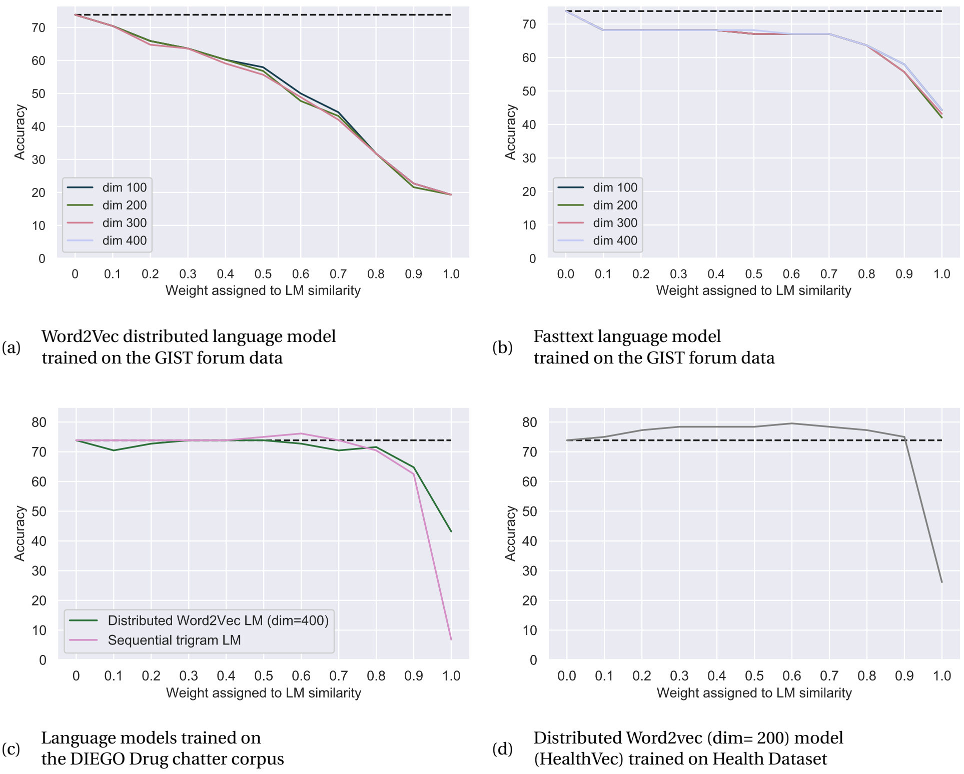
Figure 2.5: Correction accuracy of spelling mistakes with additional POS tag filters. Dotted lines indicate the best accuracy with relative edit distance alone.

distributed models of the data, we find that the higher the weight assigned to the language model similarity, the more the accuracy drops. This inverse correlation is significant and almost equal to -1 for all dimensionalities ($p < 0.000001$) (see Figure 2.6a and 2.6b). Our data is possibly too sparse to place contextual constraints on the correction candidates or to employ language model similarity in this manner. It is also too small for building a sequential trigram model [327].

Alternatively, we can employ more generic language models based on medical social media, but not specific to a particular disease domain. We find that a distributed language model based on a collection of health-related tweets, the DIEGO Drug chatter corpus [263], does not manage to improve correction accuracy (see Figure 2.6c). Nevertheless, a sequential trigram model based on this same Twitter corpus does improve correction accuracy with 2.2% point to 76.1% at a weight of 0.6 (see Figure 2.6c). The weight assigned to the probability of a trigram with the correction in place of the error is positively correlated ($r = 0.58$) with the correction accuracy. However, the HealthVec distributed language model can improve the correction accuracy up to 79.5% at a weight of 0.6 (see Figure 2.6d). Overall, its assigned weight is also positively correlated ($r = 0.63$) with the correction accuracy. Table 2.10 shows that adding the HealthVec model mostly improves accuracy for non-medical errors (e.g., 'explane') and for medical errors for which it is difficult to determine whether they should be singular or plural (e.g., 'ovarie and surgerys'). One medical term (i.e., 'surgerys') is no longer corrected accurately. We opt to employ this weighted method due to its higher overall accuracy, but one could opt to not include the HealthVec model depending on the importance of non-medical terms for the downstream task.

2.4.3. SPELLING MISTAKE DETECTION

A grid search results in an optimal similarity score threshold of 0.76. As higher similarity scores indicate that tokens are more dissimilar, this means that if the best correction candidate has a higher similarity score than this threshold, the token is not corrected (see Figure 2.7). This combination attains the maximum $F_{0.5}$ score for 8 of 10 folds. For the other two folds, 0.74 was optimal. See Figure 2.7 for the tuned decision process. On the



(a) Word2Vec distributed language model trained on the GIST forum data

(b) Fasttext language model trained on the GIST forum data

(c) Language models trained on the DIEGO Drug chatter corpus

(d) Distributed Word2vec (dim= 200) model (HealthVec) trained on Health Dataset

Figure 2.6: Correction accuracy of spelling mistakes with additional weighted language model (LM) similarity. Weight of the LM similarity is the inverse of the weight of the relative edit distance. Dotted line indicates the best accuracy with relative edit distance alone.

Table 2.10: Changes in corrections when HealthVec is added (weight = 0.6) to the relative edit distance (weight = 0.4) with $\theta = 9$. LM = language model.

	Error	Correct word	Correction	
			Without LM	With LM
Improved	alse	else	false	else
	lm	im	am	im
	esle	else	resolve	else
	explane	explain	explained	explain
	ovarie	ovary	ovary	ovaries
	surgerys	surgeries	surgeries	surgery
Missed	surgerys	surgery	surgery	surgury

test set, our method attains a significantly higher precision ($p < 0.0001$) and $F_{0.5}$ score ($p < 0.0001$) than all other detection methods (see Table 2.11). Our method does attain a slightly lower recall than dictionary-based methods, although its recall is very high at 0.91. Adding concatenation correction to our method improves recall and precision by 0.05 and 0.01, respectively. See Table 2.12 for some examples of errors made by our decision process and the corrections our method will output.

Although the recall of generic dictionaries is maximal at 1.0, their precision is low (0.11 and 0.26). Both are logical: The high recall is a result of dictionary-based methods classifying all terms *not* included in the dictionary as mistakes, which will include all non-word errors, whereas the low precision is a result of the misclassification of correct domain-specific terms that are not included in the dictionary. Aspell outperforms CELEX due to its higher coverage of relevant words such as 'oncologist', 'metastases' and 'facebook'. Microsoft Word and TISC perform the worst overall: their precision is low but they also have a lower recall than both dictionary-based methods and our method.

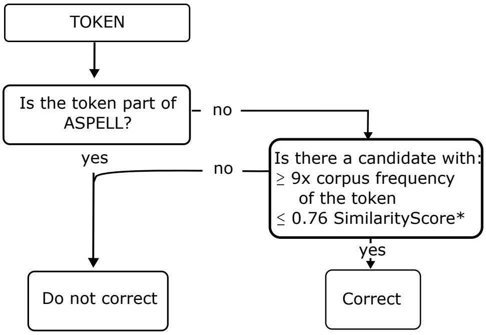

TOKEN
Is the token part of ASPELL?
no
yes
no
Is there a candidate with:
≥ 9x corpus frequency of the token
≤ 0.76 SimilarityScore*
yes
Do not correct
Correct

Figure 2.7: Decision process. *SimilarityScore = 0.6 * LM similarity + 0.4 * RE

Table 2.11: Results for mistake detection methods on the test set

Method	Mistakes found	Recall	Precision	$F_{0.5}$	F_1
CELEX	395	**1.0**	0.11	0.13	0.20
Aspell dictionary	163	**1.0**	0.26	0.31	0.42
TISC	270	0.74	0.12	0.14	0.21
Microsoft word	395	0.88	0.10	0.12	0.18
Our method (RE = 0.76)	90	0.91	0.46	0.51	0.61
Our method (RE= 0.76) + ConcatCorrection	92	0.96	**0.47**	**0.52**	**0.63**

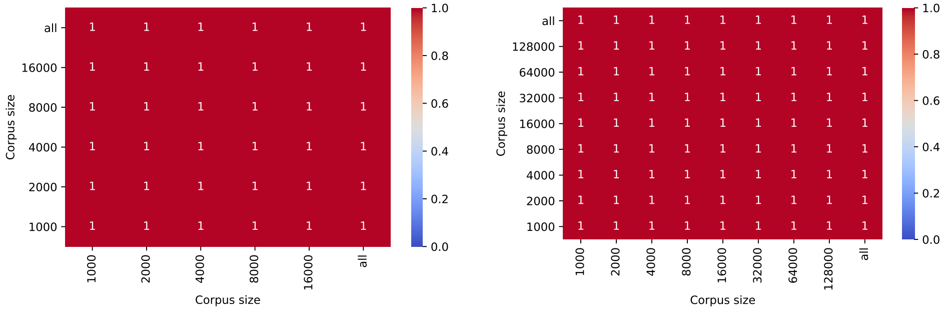
2.4.4. IMPACT OF CORPUS SIZE

Despite the fact that a relative corpus frequency threshold is more robust to different corpus sizes than an absolute one, it is likely that the ratio between tokens and their corrections will vary if the corpus size becomes smaller. Thus, we investigated to what extent the multiplication factor of 9 would be robust to such ratio changes.

Table 2.12: Examples of false positives and negatives of our error detection method.

	Mistakes (their corrections with our method)			
False positives	intolerances (intolerant)	resected (removed)	reflux (really)	condroma (syndrome)
False negatives	istological (histological)	vechile (vehicle)		

Figure 2.8 shows that our threshold is highly robust to corpus size with maximal Jaccard similarity (1.0) for all comparisons. Figure 2.9 demonstrates this with an example of one common ('gleevac') and one uncommon misspelling ('gllevec') for the medication Gleevec. The corpus frequency for each misspelling relative to the corpus size is shown with unbroken lines. The minimum corpus frequency threshold for correction candidates of each misspelling is indicated with dotted lines of the same color for the range of corpus sizes. Irrespective of the corpus size, the correct variant 'gleevec' (the purple line) remains above the minimum corpus frequency (i.e., the dotted lines) for the complete range of corpus sizes.



(a) GIST forum of 36,277 posts

(b) Reddit forum of 274,532 posts

Figure 2.8: Stability of error detection in 1000 posts with varying corpus size

2.4.5. IMPACT OF THE DEGREE OF NOISINESS OF THE DATA

As our method was designed on data with few errors (< 1% in our sample), we investigate to what extent our method can generalize to more noisy user-generated data using simulated data sets with varying proportions of misspellings. As can be seen in Figure 2.10a and 2.10b, correction accuracy is either stable or increases when the level of noise increases from 1 to 8%, whereas it appears to diminish at a noise level of 16%. As relative Levenshtein distance does not depend on the noise in the corpus, this possibly indicates that at 16% noise the corpus is affected to the degree that the frequency of correct counterparts of errors often drops below the θ of 9 times the frequency of the error. This is not surprising: due to the equal probability that each word has of being transformed into a mistake, increasingly more words necessary for correction are transformed into

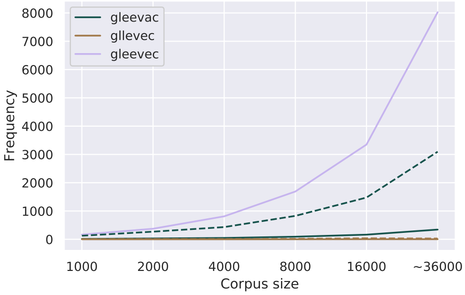
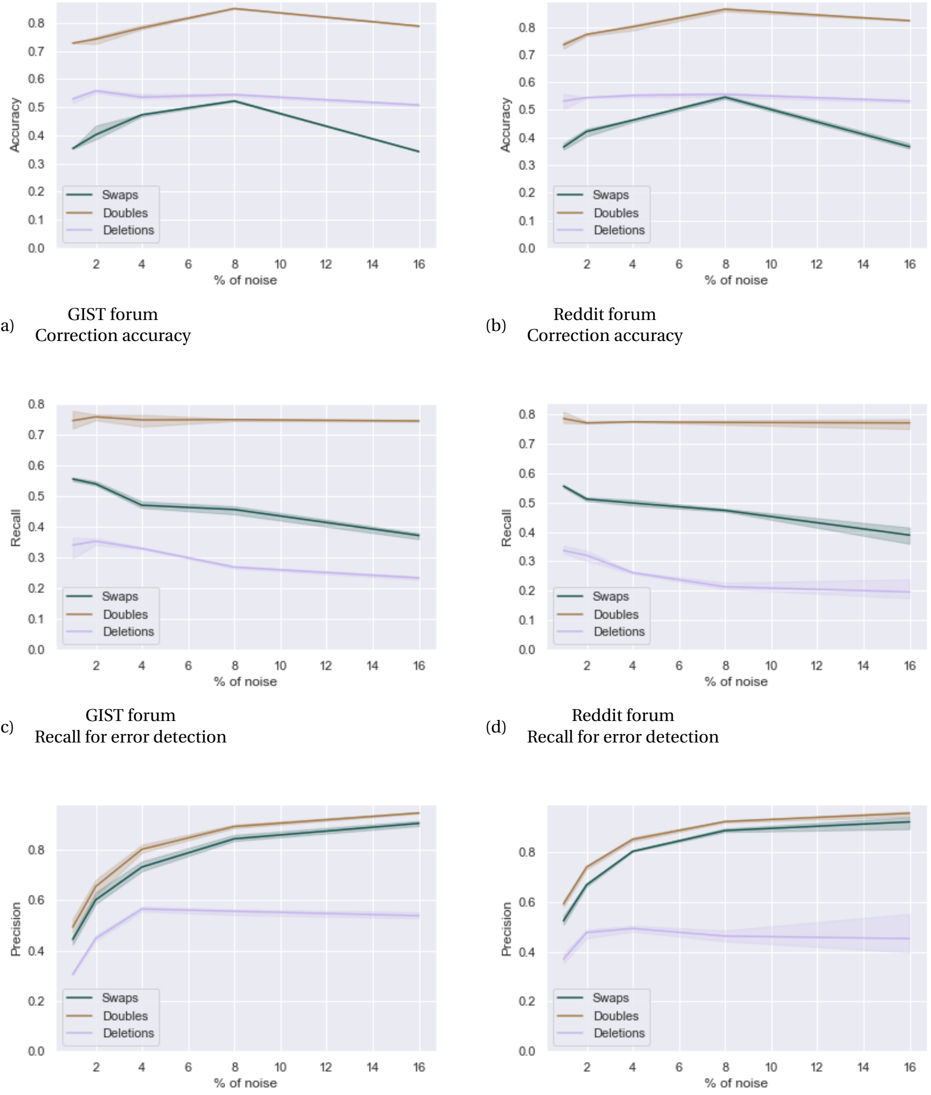


Figure 2.9: Corpus frequency of one uncommon and one common misspelling of the medication Gleevec in the GIST forum with increasing corpus size. The dotted line indicates the corpus frequency threshold for correction candidates for each misspelling.

errors. However, no conclusions can be drawn about the exact turning point, as we did not measure the impact of noise levels between 8 and 16%. If necessary, re-tuning of the threshold on a more noisy corpus may resolve this issue.

Except for errors due to doubling of letters, the absolute correction accuracy is far lower than on our real-world data set (79.5%). We believe there may be two reasons for this: firstly, users are more likely to misspell medical terms than other words [352] and thus this random distribution is unrealistic. Such medical terms are likely to be longer than the average word in social media text. Indeed, we find that in our real-world sample of 1000 posts from the GIST forum the 109 non-word errors are significantly longer than average ($p < 1e{-}22$) according to a Mann Whitney U test: The errors have a mean character length of 6.8 compared to an overall average of 4.2 characters. Since deletions or swaps in shorter words lead to more ambiguous words (e.g., ‘the’ to ‘te’) or even other real words (e.g., ‘the’ to ‘he’), this will lower the overall correction accuracy of methods designed to correct non-word medical errors. The second reason ties into this: these artificial data sets do not allow for differentiation between real word and non-word errors and thus are not suited to evaluating absolute non-word error correction. Nonetheless, although absolute accuracy on synthetic data may thus not be a reliable indicator, the relative accuracy at different noise levels does provide a good indication for the impact of the level of noise in the data on the efficacy of our method.

Regarding the detection of errors, recall appears to drop as the level of noise increases for swaps and deletions and remains roughly constant for errors due to doubling of characters (i.e., doubles) (see Figure 2.10c and 2.10d). In contrast, precision increases with increasing noise for swaps and doubles and remains mostly stable for deletions (see Figure 2.10e and 2.10f). These results may indicate that the relative frequency ratios of false positives to their predicted corrections are more frequently close to the detection threshold (θ) of 9 than those of true positives. As an artificial increase in noise by a certain percentage (e.g., 4%) will cause the frequency of correct words to drop by approximately that percentage due to random chance selection of words to transform into errors, increasing noise will lead to a slight drop in the ratio between a token and its predicted correction. If the ratio was far larger than 9, this does not alter the outcome.



(a) GIST forum
Correction accuracy

(b) Reddit forum
Correction accuracy

(c) GIST forum
Recall for error detection

(d) Reddit forum
Recall for error detection

(e) GIST forum
Precision for error detection

(f) Reddit forum
Precision for error detection

Figure 2.10: Impact of degree of noisiness of the data (1,2,4,8 and 16% noise) on the detection (c-f) and correction accuracy (a-b) of three types of spelling errors (deletions of a single letter, doubling of a single letter and swaps of adjacent letters) in two cancer-related forums. The lines indicate the mean result while the band indicates the variance in results over three runs.

However, if the ratio was only slightly above 9, then it is liable to dropping below the detection threshold when the noise is increased. In that case, the token will no longer be marked as an error. Thus, if false positives more frequently have ratios slightly above 9 than true positives do, this could explain the increase in precision.

To investigate this idea, we consider swaps in the GIST forum at different levels of noise. It appears that indeed false positives have a higher % of ratios between 9 and 10 than true mistakes at lower noise levels (2,4 and 8%) across all random seeds. This flips for 16%: false positives now have a lower percentage of ratios liable to dropping below the θ of 9 than true positives. Thus, possibly false positives that were 'at risk' for dropping below the required θ have done so. This increased precision does come at a cost: some errors will also have ratios close to 9 leading to a drop in recall with increasing noise levels.

Due to the presence of common errors, the impact of noise might be less pronounced for real data. Although the artificial data does contain common errors (e.g., 'wtih' (218x)), their frequency depends on the frequency of the word of origin (e.g., 'with' (9635x)) because each word has an equal, random probability of being altered. Consequently, their ratio will be much higher and they will be easier to detect than real common errors. Moreover, absolute precision and recall on synthetic data may not be transferable. Overall relative trends, however, do provide an first indication for the generalisability of our method to noisier data sets. Further experimentation with noisier, annotated real world data will be necessary to assess the true effect of noise on our error detection.

For both error correction and detection, results are consistent across the two forums and variance of the results is low except at tail end (16%). This can be explained by the random assignment of transformations for each run: depending on which words are randomly transformed in a certain run, the frequency of certain correct words may either fall below the θ of 9 or not.

2.4.6. EFFECT ON OOV RATE

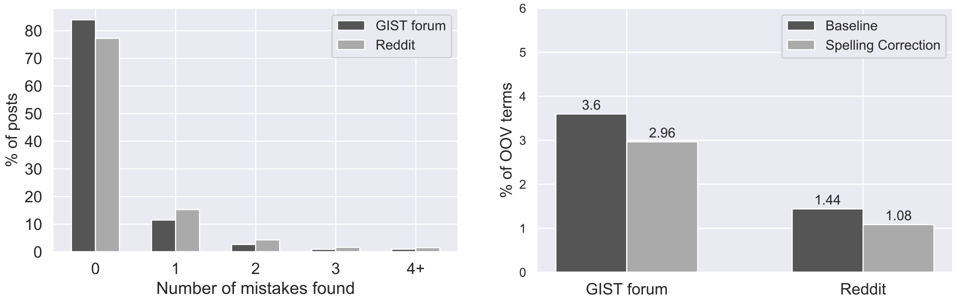
The reduction in out-of-vocabulary (OOV) terms is higher for the GIST (0.64%) than for the Reddit forum (0.36%) (See Figure 2.11b). As expected, it appears that in-vocabulary terms are occasionally replaced with out-of-vocabulary terms, as the percentage of altered words is higher than the reduction in OOV (0.72% vs 0.64% for the GIST and 0.50% vs 0.36% for the Reddit forum). The vast majority of the posts do not contain any mistakes and of the posts with mistakes, the majority have only one (see Figure 2.11a). Thus, it appears that the spelling mistakes are not caused by a select group of individuals that are poor at spelling, but by various forum users making the occasional mistake.

Interestingly, the prior OOV count of the GIST forum is more than double that of the sub-reddit on cancer. This could be explained by the more specific nature of the forum: it may contain more words that are excluded from the dictionary, despite the fact that the dictionary is tailored to the cancer domain. This again underscores the limitations of dictionary-based methods.

Many of the most frequent corrections made in the GIST forum are medical terms (e.g., gleevec, oncologists, tumors). Similarly, the most frequent mistakes found in this forum are common misspellings of medical terms (e.g., gleevac and gleevic) (see Figure 2.12a). It appears that for common medical corrections, there are often various less commonly occurring misspellings per term since their misspelt equivalents do not show up amongst the most common mistakes. We also found that our method normalizes variants of medical terms to the more prevalent one (e.g., reoccurrence to recurrence). Thus, although the overall reduction in OOV-terms may seem minor, our approach appears to target medical concepts, which are highly relevant for knowledge extraction tasks. In addition, our method incorrectly alters plural to singular variants (e.g., gists to gist), probably due to their higher prevalence in the data. Additionally, due to the addition of the distributed language model, prevalent terms can be replaced by their synonyms as 'corrections' (e.g., resected to removed). Fortunately, the resulting information loss will be minimal for medical downstream tasks.

In the sub-reddit on cancer, frequent corrections include medical terms (e.g., chemotherapy, medication and hospital), normalization from plural to singular (e.g., wifes to wife) but also both incorrect alterations of slang (e.g., gon to got) and of medical terms (e.g., immunotherapy) (see Figure 2.12b). Additionally, the change from didn to did is problematic due to the loss of the negation. Our method thus appears to work less well for more general fora.

Nonetheless, when we consider the 50 most frequent remaining OOV terms, only a small proportion of them are non-word spelling errors, although slang words could arguably also be part of this category (see Table 2.13 for examples). A significant portion consists of real words not present in the specialized dictionary. Importantly, also some drug names and medical slang (e.g., 'scanxiety' or anxiety about being scanned) are considered OOV. Since they can be essential for downstream tasks, it is promising that they have not been altered by our method.



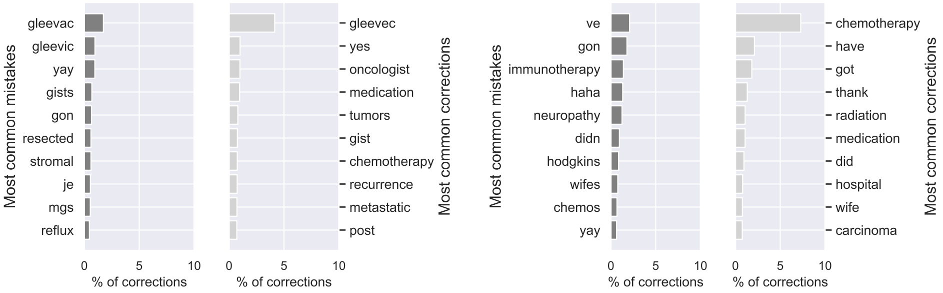
(a) Distribution of found mistakes across posts

(b) Change in out-of-vocabulary terms

Figure 2.11: Internal validation on two cancer forums

2.4.7. External validation

As can be seen in Table 2.14, normalization leads to a significant change in the F_1 score for two of the six classification tasks ($p = 0.0096$ and $p = 0.0044$). For the Twitter Health corpus, this change is mostly likely driven by a significant increase in recall ($p = 0.0040$), whereas for the detection of flu infection tweets (Task4 SMM4H2019) it is the precision that is increased significantly ($p = 0.0064$). In general, these changes are of the same order of magnitude as those made by the normalization pipeline of Sarker [261]. Although the



(a) GIST forum

(b) Reddit forum on cancer

Figure 2.12: Most frequent mistakes and corrections in two cancer forums

Table 2.13: Manual error analysis of 50 most frequent OOV terms after spelling detection

	GIST	Example	Reddit	Example
Real word	33	unpredictable, internet	42	misdiagnosed, website
Spelling mistake	5	side-effects, wildtype, copay	2	side-effects, inpatient
Abbreviation	2	mos, wk	3	aka
Slang	6	scanxiety, gister	1	rad
Drug name	2	stivarga, mastinib	1	ativan
Not English	2	que, moi	-	
TOTAL	50		50	

overall classification accuracy on Task 1 of the SMM4H workshop is low, this is in line with the low F_1 score (0.522) of the best performing system on a comparable task in 2018 [335].

Especially the expansion of contractions and the splitting of hash tags (e.g., ‘#flushot’ to ’#flu shot’) appear to impact the classification outcome. In contrast, neither the goal of the task, the relative amount of corrections nor the initial result seem to correlate with the change in F_1 score. The lack of a correlation between the amount of alterations and the change in F_1 score may be explained by the weak reliance of classification tasks on individual terms. Unlike in Sarker [261], the improvements also do not seem to increase with the size of the data. This is logical, as we do not rely on training data. The imbalance of the data may be associated with the change in accuracy to some extent: the two most balanced data sets show the largest increase (see Table 2.3). Further experiments would be necessary to elucidate if this is truly the case.

On generic social media text, our method performs only slightly worse than the state-of-the-art methods (see Table 2.15). We did not need to use the training data, as our method is unsupervised. For comparison, our method attains a F_1 of 0.726, a precision of 0.728, and a recall of 0.726 on the W-NUT training data.

Error analysis reveals that 46 of the 100 most frequent remaining errors are words that should not have been altered according to the W-NUT annotation (see Table 2.16). Yet, in fact, these words are often slang that our method expanded correctly (e.g., info to information). It is thus debatable whether these are errors. Of the remainder, 33 are either

Table 2.14: Mean classification accuracy before and after normalization for six health-related classification tasks. Only the results for the best performing classifier per data set are reported. $**$ indicates $p<0.005$; $*$ indicates $p<0.01$; † indicates absolute change

		F1		Recall		Precision	
Data set	Words altered	Pre	Δ †	Pre	Δ †	Pre	Δ †
Task1 SMM4H 2019	1.53%	0.410	-0.0007	0.373	+0.014	0.470	-0.025
Task4 SMM4H 2019 Flu Vaccination	0.50%	0.780	+0.006	0.834	+0.008	0.733	+0.005
Flu Vaccination Tweets	0.50%	0.939	+0.002	0.935	+0.004	0.943	+0.0004
Twitter Health	0.71%	0.702	+0.016*	0.657	+0.028*	0.756	-0.0009
Task4 SMM4H 2019 Flu Infection	0.57%	0.784	+0.012**	0.842	+0.013	0.735	+0.019**
Zika Conspiracy	0.36%	0.822	-0.005	0.817	+0.012	0.835	-0.021

uncorrected abbreviations or slang terms. This may partially be explained by the fact that the slang usage of medical forum users differs from the general Twitter population. Lastly, 16 of these 100 can be considered non-word errors that were missed by our method and another 4 are errors that were correctly detected but corrected inaccurately.

	F_1	Precision	Recall
MoNoise [318]	**0.864**	**0.934**	0.803
Sarker's method [261]	0.836	0.880	0.796
IHS_RD [292]	0.827	0.847	**0.808**
USZEGED [31]	0.805	0.861	0.756
BEKLI [24]	0.757	0.774	0.742
LYSGROUP [89]	0.531	0.459	0.630
Our method	0.743	0.734	0.753

Table 2.15: Results for unconstrained systems of ACL W-NUT 2015

2.5. DISCUSSION

The state-of-the-art normalization method for generic social media [261] performs poorly for medical social media with a spelling correction accuracy of only 19.3%. It is significantly outperformed by all edit-based methods regardless of whether the correction candidates are derived from a specialized vocabulary or the data itself. The highest correction accuracy (73.9%) is attained by unweighted relative edit distance combined with the constraint that correction candidates need to be at least 9 times more frequent than the error. This accuracy is further increased by 5.6% point to 79.5% when it is combined with model similarity based on the HealthVec language model. Our preceding decision process is capable of identifying mistakes for subsequent correction with a $F_{0.5}$ of 0.52 and a significantly higher precision than all other methods while retaining a very high recall of 0.96. Additionally, it is almost completely independent of the size of the corpus for the two cancer-related forums, which is promising for its usage in other even smaller,

Table 2.16: Manual analysis of 100 most frequent errors in W-NUT. *also considered non-word mistakes

Type of error	Freq.	Example	Our correction	W-NUT annotation
Should not have been altered	46	info, kinda	information, kind of	info, kinda
Abbreviation not or incorrectly expanded	19	smh	smh	shaking my head
Uncorrected slang	14	esp	esp	especially
Missed concatenation error*	6	incase	incase	in case
Missed apostrophe*	5	youre	youre	you're
Wrong correction	4	u	your	your
Missed split mistake*	3	i g g y	i g g y	iggy
Missed non-word spelling mistake	2	limites	limites	limits
American English	1	realise	realize	realise
TOTAL	100			

domain-specific data sets. Our method can also function well for more noisy corpora up to a noise level of 8% (i.e., 1 error in every 12.5 words).

In the two cancer forums that we used for evaluation, the spelling correction reduces OOV-terms by 0.64% point and 0.36% point. Although the reduction may seem minor, relevant medical terms appear to be targeted and, additionally, many of the remaining OOV-terms are not spelling errors but rather real words, slang, names, and abbreviations. Furthermore, our method was designed to be conservative and to focus on precision to mitigate one of the major challenges of correcting errors in domain-specific data: the loss of information due to the 'correction' of correct domain-specific terms.

Our method also significantly improves the classification accuracy on two tasks, although the absolute change is marginal. On the one hand, this could be because classification tasks do not rely strongly on individual terms. On the other hand, it may be explained by our use of only unigrams as features. Feature extraction would likely also benefit from normalization and could possibly increase performance differences. Further experimentation is required to further assess the full effect of lexical normalization in downstream tasks.

As named entity recognition (NER) tasks rely more strongly on individual terms, we speculate that our method will have a larger impact on such tasks. Unfortunately, NER benchmarks for health-related social media are limited. We have investigated three relevant NER tasks that were publicly available: CADEC [151], ADRMiner [217], and the ADR extraction task of the SMM4H 2019. For all three tasks, extracted concepts could be matched exactly to the forum posts, thus negating the potential benefit of normalization. The exact matching can perhaps be explained by the fact that data collection and extraction from noisy text sources such as social media typically rely on keyword-based searching [264].

Our study has a number of limitations. Firstly, the use of OOV-terms as a proxy for the quality of the data relies heavily on the vocabulary that is chosen and, moreover, does not allow for differentiation between correct and incorrect substitutions. Secondly, our method is currently targeted specifically at correcting non-word errors and is therefore unable to correct real word errors. Thirdly, the evaluation data set for developing our method is small: a larger evaluation data set would allow for more rigorous testing. Nonetheless, as far as we are aware, our corpora are the first for evaluating mistake detection and correction in a medical patient forum. We welcome comparable data sets sourced from various patient communities for further refinement and testing of our method.

2.6. CONCLUSIONS AND FUTURE WORK

To what extent can corpus-driven spelling correction reduce the out-of-vocabulary rate in medical social media text? Our corpus-driven spelling correction reduces the OOV rate by 0.64% point and 0.36% point in the two cancer-related medical forums we used for evaluation. More importantly, relevant medical terms appear to be targeted.

To what extent can corpus-driven spelling correction improve accuracy of health-related classification tasks with social media text? Our corpus-driven method could significantly improve the classification accuracy on two of the six tasks. This is driven by a significant increase in precision for one and by a significant increase in recall for the second.

In conclusion, our data-driven, unsupervised spelling correction method can improve the quality of text data from medical forum posts. We have demonstrated the success of our method on data from two cancer-related forums. The automatic spelling corrections significantly improve the F_1 score for two of the six external classification tasks that involve medical social media data. Our method can also be useful for user-generated content in other highly specific and noisy domains, which contain many OOV terms compared to available dictionaries. Future work will include extending the pipeline with modules for named entity recognition, automated relation annotation and concept normalization. Another possible avenue for future work could be to determine whether a word is or is not from the domain at hand (the medical domain in our case) prior to normalization and apply different normalization techniques in either case. Furthermore, despite a lack of domain-specific, noisy corpora for training character-level language models, it would be interesting to investigate to what extent our spelling correction can improve classification accuracy using character-level language models pretrained on other source domains.