Lexical Normalization of User-Generated Medical Forum Data

Anne Dirkson, Suzan Verberne & Wessel Kraaij
LIACS, Leiden University
Niels Bohrweg 1, Leiden, the Netherlands
{a.r.dirkson, s.verberne, w.kraaij}@liacs.leidenuniv.nl

Abstract

In the medical domain, user-generated social media text is increasingly used as a valuable complementary knowledge source to scientific medical literature. The extraction of this knowledge is complicated by colloquial language use and misspellings. Yet, lexical normalization of such data has not been addressed properly. This paper presents an unsupervised, data-driven spelling correction module for medical social media. Our method outperforms state-of-the-art spelling correction and can detect mistakes with an $F_{0.5}$ of 0.888. Additionally, we present a novel corpus for spelling mistake detection and correction on a medical patient forum.

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) (Gonzalez-Hernandez et al., 2017). 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 (Sarker et al., 2016) and the detection of adverse drug responses (Sarker et al., 2015). However, the extraction of this knowledge is complicated by non-standard and colloquial language use, typographical errors, phonetic substitutions, and misspellings (Clark and Araki, 2011; Sarker, 2017; Park et al., 2015). Thus, social media text is generally noisy and this is only aggravated by the complex medical domain (Gonzalez-Hernandez et al., 2017).

Despite these challenges, text normalization for medical social media has not been explored thoroughly. Medical lexical normalization methods (i.e. abbreviation expansion (Mowery et al., 2016) and spelling correction (Lai et al., 2015; Patrick et al., 2010)) have mostly been developed for clinical records or notes, as these also contain an abundance of domain-specific abbreviations and misspellings. However, social media text presents distinct challenges, such as colloquial language use, (Gonzalez-Hernandez et al., 2017; Sarker, 2017) that cannot be tackled with these methods.

The most comprehensive benchmark for general-domain social media text normalization is the ACL W-NUT 2015 shared task[1] (Baldwin et al., 2015). The current state-of-the-art system for this task is a modular pipeline with a hybrid approach to spelling, developed by Sarker (2017). Their pipeline also includes a customizable back-end module for domain-specific normalization. However, this back-end module relies, on the one hand, on a standard dictionary supplemented manually with domain-specific terms to detect mistakes and, on the other hand, on a language model of generic Twitter data to correct these mistakes. For domains that have many out-of-vocabulary (OOV) terms compared to the available dictionaries and language models, 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. Consequently, hand-made lexicons may get outdated (Gonzalez-Hernandez et al., 2017). In this paper, we present an alternative: a corpus-driven spelling correction approach. 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?

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

Proceedings of the 4th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 11–20
Florence, Italy, August 2, 2019.

2. To what extent can our corpus-driven spelling correction improve 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 and (2) the first corpus for evaluating mistake detection and correction in a medical patient forum.[2]

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 'correction' of already correct domain-specific terms. We hypothesize that a dictionary-based method is able to retrieve more mistakes than a data-driven method, because all terms *not* included in the dictionary are classified as mistakes, which will probably include all non-word errors. However, we also expect that a dictionary-based method will misclassify more correct terms as mistakes, because any domain-specific terms not present in the dictionary will be classified incorrectly.

2 Related work

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 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. a 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 (Zeng and Tse, 2006). Another complication is the frequent misspellings of key medical terms, as medical terms are typically difficult to spell (Zhou et al., 2015). 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.

Lexical normalization of generic social media In earlier research, text normalization for social media was mostly unsupervised or semi-supervised e.g. (Han et al., 2012) due to a lack of annotated data. These methods often pre-selected and ranked correction candidates based on phonetic or lexical string similarity (Han et al., 2012, 2013). Han et al. (2013) additionally used a trigram language model trained 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 (Han et al., 2012, 2013). The opposite is true for modern supervised methods, which rely on training data but not on dictionaries. 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 (Jin, 2015). The second and third best performing methods were also supervised and used deep learning to detect and correct mistakes (Leeman-Munk et al., 2015; Min and Mott, 2015) (for more detail on W-NUT systems see Baldwin et al. (2015)). Since 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.

Additionally, recent approaches often make use of language models, which require a large corpus of comparable text from the same genre and domain (Sarker, 2017). This is however 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 (Sarker, 2017) due to their ability to capture the context of words and to handle the dynamic nature of language.

3 Data

Medical forum data For evaluating spelling correction methods, we use an international pa-

[2]The corpus is available on github https://github.com/AnneDirkson

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

Table 1: Raw data without punctuation. IQR: Interquartile range

tient forum for patients with Gastrointestinal Stromal Tumor (GIST). It is moderated by GIST Support International (GSI). This data set was donated to us by GSI in 2015. We use a second cancer-related forum to assess generalisability of our methods: a sub-reddit community on cancer, dating from 16/09/2009 until 02/07/2018.[3] It was scraped using the Pushshift Reddit API.[4] The data was collected by looping over the timestamps in the data. This second forum is around 4x larger than the first in terms of tokens (See Table 1).

Annotated data Spelling mistakes were annotated for 500 randomly selected posts from the GIST data. Real word errors and split or concatenation errors were not included, because we are not interested in syntactic or semantic errors (Kukich, 1992). In addition, we considered each word independent of its content, because word bigrams or trigrams are sparse in the small forum collections (Verberne, 2002). Each token was classified as a mistake (1) or not (0) by the first author. A second annotator checked if any of the mistakes were false positives. 53 unique mistakes were found: Their corrections were annotated individually by two annotators. Annotators were provided with the complete post in order 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. These 53 mistakes and their corrections form the test set for evaluating spelling correction methods.[5] As far as we are aware, no other spelling error corpora for this domain are publicly available.

In order to tune various thresholds for the detection of spelling mistakes, we split these 500 posts into two sets of 250 posts: a development and a test set. The development set contained 23 mistakes supplemented with a tenfold of randomly selected correct words (230) with the same word length distribution. The development set was split in a stratified manner into 10 folds for cross-validation. The test set contained 32 unique non-word errors [6], equal to 0.37% of the tokens, supplemented with a tenfold of randomly selected correct words with the same word length distribution.[7]

Spelling error frequency corpus 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 corpus.

Specialized vocabulary for cancer forums To be able to calculate the number of out-of-vocabulary terms in two cancer forums, a specialized vocabulary was created by merging the standard English lexicon CELEX (Burnage et al., 1990) (73,452 tokens), the NCI Dictionary of Cancer Terms (National Cancer Institute) (6,038 tokens), the generic and commercial drug names from the RxNorm (National Library of Medicine (US)) (3,837 tokens), the ADR lexicon used by Nikfarjam et al. (2015) (30,846 tokens) and our in-house domain-specific abbreviation expansions (DSAE) (42 tokens) (see Preprocessing 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).

Data sets for external validation We obtained six public classification data sets that use health-related social media data. They were retrieved from the data repository of Dredze[9] and the shared tasks of Social Media Mining 4 Health workshop (SMM4H) 2019[10]. The data sets sizes range from 588 to 16,141 posts (see Table 2).

[3] www.reddit.com/r/cancer

[4] https://github.com/pushshift/api

[5] Corpora and code are available on github https://github.com/AnneDirkson

[6] Two errors overlapped between the sets

[7] Due to a limited number of words of length 17, 311 instead of 320 words were added

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

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

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

Data set	Task	Size	Positive (%)	Negative (%)
Task 1 SMM4H 2019*	Presence adverse drug reaction	16,141	8.7	91.3
Task 4 SMM4H 2019* Flu vaccine	Personal health mention of flu vaccination	6,738	28.3	71.7
Flu Vaccination Tweets (Huang et al., 2017)	Relevance to topic flu vaccination	3,798	26.4	73.6
Twitter Health (Paul and Dredze, 2009)	Relevance to health	2,598	40.1	59.9
Task4 SMM4H 2019* Flu infection	Personal health mention of having flu	1,034	54.4	45.6
Zika Conspiracy Tweets (Dredze et al., 2016)	Contains pseudo-scientific information	588	25.9	74.1

Table 2: Six classification data sets of health-related Twitter data. *SMM4H: Social Media Mining 4 Health workshop

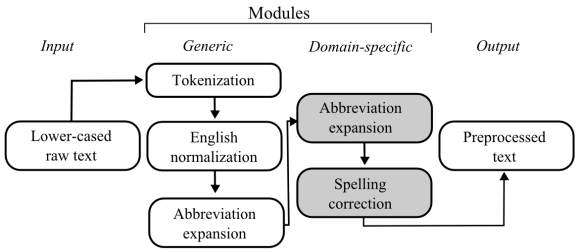

Figure 1: Sequential processing pipeline

4 Methods

Preprocessing To protect the privacy of users, in-text person names were replaced as much as possible using a combination of the NLTK names corpus and part-of-speech tags (NNP and NNPS). Additionally, URLs and email addresses were replaced by the strings -url- and -email- using regular expressions. Furthermore, text was lowercased and tokenized using NLTK. The first modules of the normalization pipeline of Sarker (2017) were employed: converting British to American English and normalizing generic abbreviations (see Figure 1). Some forum-specific additions were made: Gleevec (British variant: Glivec) was included in the British-American spelling conversion and one generic abbreviation expansion that clashed with a domain-specific one was substituted (i.e. 'temp' defined as *temperature* instead of *temporary*). Moreover, the abbreviations dictionary by Sarker (2017) was lowercased. Lastly, domain-specific abbreviations were expanded with a lexicon of 42 non-ambiguous abbreviations, generated based on 500 randomly selected posts from the GIST forum and annotated by a domain expert and the first author. [11]

[11]This lexicon is shared on github https://github.com/AnneDirkson

Spelling correction We used the method by Sarker (2017) 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 paper, we used only the DIEGO LAB Drug Chatter Corpus (Sarker and Gonzalez, 2017a), 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 (2005). 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 costs for an edit, the following transformation of the log frequencies was used: Weight Edit Distance = $\frac{1}{1+log(frequency)}$.

Spelling mistake detection We manually constructed a decision process, inspired by the work by Beeksma et al. (2019), for detecting spelling mistakes (See Figure 2). The decision process uses the corpus frequency relative to that of the token and the similarity to the token. The underlying idea is that if a word is either common within the domain-specific language or there is no simi-

lar enough candidate available, it is unlikely to be a mistake. A relative threshold enables us to capture more common mistakes.

To ensure generalisability, we opted for an unsupervised, data-driven method that does not rely on the construction of a specialized vocabulary. Candidates are considered in order of frequency. Of the candidates with the highest similarity score, the first is selected. The spelling correction ignores numbers and punctuation.

To optimize the decision process, a 10-fold cross validation grid search was conducted with a grid of 2 to 10 (steps of 1) for the minimum multiplication factor of the corpus frequency and a grid of 0.05 to 0.15 (steps of 0.01) for the minimum similarity. The choice of grid was based on previous work by Walasek (2016) and Beeksma et al. (2019). 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.

Spelling correction candidates For evaluating the mistake detection process, spelling correction candidates are derived from the data itself using the corpus frequency and similarity thresholds. For internal and external validation, candidates are also derived from the data itself. However, for comparing the spelling correction methods, the words of the specialized vocabulary for cancer forums (see section 3) were used as correction candidates in order to evaluate the methods independently of the vocabulary present in the data.

Internal validation 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 is 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 CELEX may also be transformed from IV to OOV. The forum text was lemmatised prior to spelling correction. OOV analysis was done manually.

External validation Text classification was performed with default sklearn classifiers: Stochastic Gradient Descent (SGD), Multinomial Naive Bayes (MNB) and Linear Support Vector Machine (SVC). Uni-grams were used as features. A 10-fold cross-validation was used to determine the average score and 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.

To evaluate our method on generic social media text, we used the test set of the ACL W-NUT 2015 task (Baldwin et al., 2015). 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. For comparison, the F_1 score on the W-NUT training data was 0.562.

	Accuracy
Sarker's method	20.8 %
TISC	24.5 %
Absolute Edit distance (AE)	56.6 %
Relative Edit distance (RE)	56.6 %
Absolute Weighted Edit distance (AWE)	54.7 %
Relative Weighted Edit distance (RWE)	**62.3%**
Upper bound	84.9%

Table 3: Accuracy of spelling correction methods

5 Results

5.1 Spelling correction

The state-of-the-art method for generic social media performed poorly on medical social media with an accuracy of only 20.8% (see Table 3). A second established data-driven approach, TISC, also performed poorly (24.5%). The best performing baseline method on our spelling corpus was Relative Weighted Edit distance (RWE) (62.3%). As eight corrections did not occur in the CELEX, the upper bound was 84.9%.

One of the reasons for the low accuracy of Sarker's method may be the absence of correct terms (e.g. gleevec) in the language model it employs. This potential complication was already highlighted by Sarker (2017) in their own paper. Similarly, the large corpus of English news texts, which TISC relies on, may not contain the right terms or may not be comparable enough as a language model to our domain-specific data set.

In contrast, the key to the success of weighted edit distance methods is likely the incorporation of probabilities for 1-edit errors. This matches the intuition that certain errors are easier to make than others. For example, someone is more likely to wrongly spell sutent as sutant than as mutant (see Table 4). Such weighted methods indirectly integrate different types of possible errors, such as typo- and orthographical errors. The relative

Mistake	gllevec	stomack	sutant
Correct	gleevec*	stomach	sutent*
Sarker's method	clever	smack	mutant
TISC	gllevec	smack	dunant
AE	**gleevec**	**stomach**	mutant
RE	**gleevec**	**stomach**	mutant
AWE	**gleevec**	smack	**sutent**
RWE	**gleevec**	**stomach**	**sutent**

Table 4: Corrections made by spelling methods. *Gleevec and Sutent are important cancer medications for GIST patients

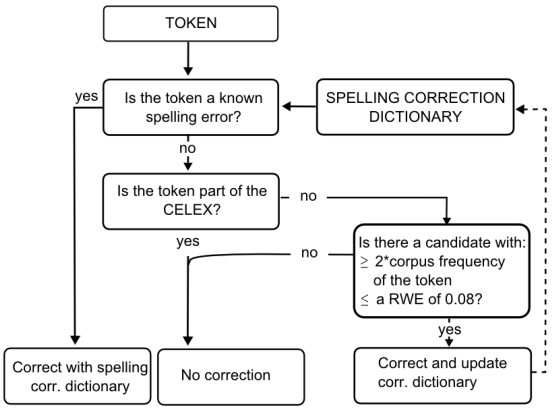

Figure 2: Decision process for spelling corrections. RWE: Relative Weighted Edit Distance

variant, as opposed to the absolute weighted edit distance, can counterbalance cheap deletions and additions, as can be seen for the mistake *stomack* (See Table 4).

5.2 Detecting spelling mistakes

The grid search results in two criteria for correction candidates: (1) a minimum of 2 times the relative corpus frequency of the token and (2) a maximum similarity score of 0.08 (see Figure 2). This combination attains the maximum $F_{0.5}$ score for all 10 folds.

On the test set, the decision process has an $F_{0.5}$ of 0.888. Its precision is high (0.90). Although the recall of a generic dictionary (i.e. CELEX) is maximal (1.0), its precision is low (0.464). This indicates, as hypothesized, that a dictionary-based method can retrieve more of the mistakes, but also will identify many correct terms as mistakes. Some examples of false positives were: 'oncologist', 'gleevec' and 'colonoscopy'. See Table 6 for some examples of errors made by our decision process.

The accuracy of the RWE method is further increased by 1.8% point by filtering the correction candidates using the preceding decision process,

	$F_{0.5}$	F_1	Recall	Precision
CELEX	0.519	0.634	**1.0**	0.464
Decision process	**0.888**	**0.871**	0.844	**0.900**

Table 5: Results for mistake detection methods on the test set

False positives	oncologists	recruiter	angiogram
False negatives	norvay	stomach	vac

Table 6: Examples of errors of the decision process

as is done in the full spelling module. The upper limit for spelling correction also increased from 84.9% to 92.5% by using candidates from the data instead of a specialized dictionary.

5.3 Effect on OOV rate

The reduction in OOV-terms was higher for the GIST (0.50%) than for the Reddit forum (0.27%) (See Figure 3). 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 (0.67% vs 0.50% for the GIST and 0.44% vs 0.27% for the Reddit forum).

Interestingly, the initial OOV count before spelling correction of the GIST forum is almost 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 it being tailored to the cancer domain. This again underscores the limitations of dictionary-based methods.

Some of the most frequent corrections made in the GIST forum data were medical terms (e.g. gleevec, scan). 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. Besides correcting mistakes in medical terms, our method also normalizes variants of medical terms (e.g. metastatic to metastasis). This is possibly a result of the corpus frequency comparison between tokens and candidates, which favors more prevalent variants.

Concerning the 50 most frequent remaining OOV terms, only a small proportion of them are in fact non-word spelling errors (e.g. 'wa'), although slang words (e.g 'ya') could arguably also be part of this category (see Table 7). A significant portion consists of real words (e.g. 'online', 'website', 'stressful') not present in the specialized dictio-

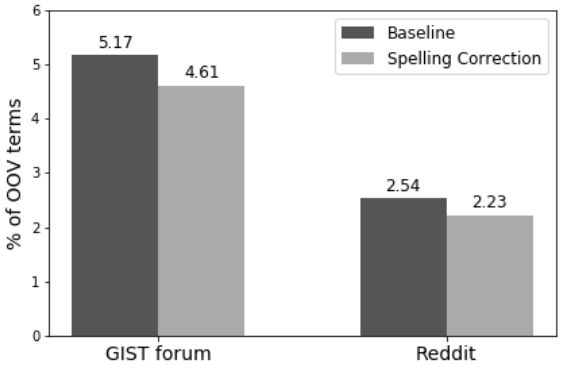

Figure 3: Percentage of OOV-terms in two cancer forums pre- and post-spelling correction.

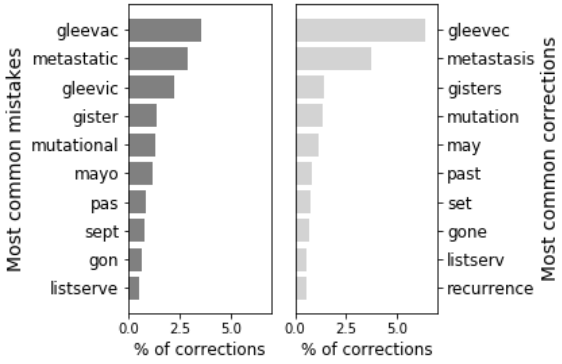

Figure 4: Most frequent mistakes and corrections in the GIST forum

nary. Upon manual inspection, the abbreviations frequently refer to treatments (e.g. 'rai'), mutation types (e.g. 'nf') or hospitals (e.g 'ucla'). Importantly, also some drug names are considered OOV (e.g. 'ativan'). Since they can be essential for downstream tasks, it is promising that they have not been altered by our method.

5.4 External evaluation

As can been seen in Table 8, the spelling correction does not lead to significant changes in the F_1 score for five of the six tasks. For the Twitter Health classification task, the improvement is significant with a p-value of 0.041 according to a paired t-test.

	GIST forum	Reddit
Spelling error	3	1
Real word	11	21
Abbreviation	14	9
Slang	6	13
Name of person or hospital	14	2
Drug name	1	4
Not English	1	0
TOTAL	50	50

Table 7: Analysis of 50 most frequent remaining OOV in two cancer forums

In general, these changes are of the same order of magnitude as those made by the normalization pipeline of Sarker (2017). Moreover, the % of alterations due to spelling correction is comparable to that of the two cancer-related forums (see Figure 3). Although the 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 the comparable task in 2018 (Weissenbacher et al., 2018).

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. Unlike in Sarker (2017), the improvements also do not seem to increase with the size of the 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). Further experiments would be necessary to elucidate if this is truly the case.

As can be seen in Table 9, our method does not perform well on generic social media text. In comparison, Sarker (2017)'s method attained state-of-the-art results with a F_1 of 0.836 on the ACL W-NUT 2015, but functioned poorly for medical social media (see Table 3). Thus, the success on one does not imply success on the other and consequently, normalisation of generic social media text and of domain-specific social media text appear different to the extent that they necessitate different approaches.

6 Discussion

Relative weighted edit distance outperforms both Sarker's method and other edit distance metrics with an accuracy of 62.3%. The accuracy is increased by a further 1.8% point if correction candidates are filtered with the criteria of the preceding decision process. This decision process is also capable of identifying mistakes with an $F_{0.5}$ of 0.888 and a high precision (0.90).

The spelling correction method led to an overall reduction in OOV-terms of 0.50% and 0.27% for two cancer-related forums. Although the reduction of OOV-terms may seem minor, relevant medical terms appear to be targeted (see Figure 4) and, additionally, many of the remaining OOV-terms are not spelling errors (see Table 7). Further-

Data set	Classifier	Prenorm F_1	Postnorm F_1	Postspell F_1	Change$^+$	% of words corrected
Task 1 SMM4H 2019	SVC	0.410	0.413	0.417	+0.006	1.1
Task 4 SMM4H 2019 Flu vaccine	MNB	0.780	0.781	0.782	+0.001	0.47
Flu Vaccination Tweets	SVC	0.939	0.938	0.941	+0.002	0.83
Twitter Health	MNB	0.702	0.708	0.713	+0.010*	0.64
Task4 SMM4H 2019 Flu infection	MNB	0.784	0.792	0.795	+0.011	0.29
Zika Conspiracy Tweets	MNB	0.822	0.818	0.811	-0.011	1.1

Table 8: Mean classification accuracy before normalization (prenorm), after normalization (postnorm) and after spelling correction (postspell) for six health-related classification tasks. Only the results for the best performing classifier per data set are reported. MNB: Multinomial Naive Bayes; SVC: Linear Support Vector Classification. $^+$Absolute change compared to prenorm.

	F_1	Precision	Recall
Sarker's method (Sarker, 2017)	**0.836**	**0.880**	0.796
IHS_RD (Supranovich and Patsepnia, 2015)	0.827	0.847	**0.808**
USZEGED (Berend and Tasnádi, 2015)	0.805	0.861	0.756
BEKLI (Beckley, 2015)	0.757	0.774	0.742
LYSGROUP (Doval Mosquera et al., 2015)	0.531	0.459	0.630
Our method	0.522	0.646	0.577

Table 9: Results for unconstrained systems of ACL W-NUT 2015

more, 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. The marginal change in task-based classification accuracy may be due to the fact that classification tasks do not rely strongly on individual terms, but on all words combined. This could also explain the lack of a correlation between the amount of alterations and the change in F_1 score. We plan to evaluate these results further by analysing both the corrections and the classification errors.

We speculate that our method will have a larger impact on named entity recognition (NER) tasks. Unfortunately, NER benchmarks for health-related social media are limited. We have investigated three relevant NER tasks that were publicly available: CADEC (Karimi et al., 2015), ADR-Miner (Nikfarjam et al., 2015), 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 (Sarker and Gonzalez, 2017b).

Our study has a number of limitations. Firstly, the use of OOV-terms as a proxy for quality of the data relies heavily on the vocabulary that is chosen and, moreover, does not allow for differentiation between correct and incorrect substitutions. Consequently, we also test whether our method can improve classification accuracy on various tasks. Secondly, our method is currently targeted specifically at correcting non-word errors and is thus is unable to correct real word errors. Thirdly, our 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.

7 Conclusion and future work

Our data-driven, unsupervised spelling correction can improve the quality of text data from medical forum posts from two cancer-related forums. Our method may also be useful for user-generated content in other highly specific and noisy domains, which contain many OOV compared to available dictionaries. Future work will include extending the pipeline with modules for named entity recognition, automated relation annotation and concept normalization.

8 Acknowledgements

We thank the SIDN fonds for financing this project and Abeed Sarker for his valuable feedback.

References

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 126–135. Association for Computational Linguistics.

Russell Beckley. 2015. Bekli: A simple approach to twitter text normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 82–86, Beijing, China. Association for Computational Linguistics.

Merijn Beeksma, Suzan Verberne, Antal van den Bosch, Iris Hendrickx, Enny Das, and Stef Groenewoud. 2019. Predicting life expectancy with a recurrent neural network. *BMC Medical Informatics and Decision Making*. To appear.

Gábor Berend and Ervin Tasnádi. 2015. Uszeged: Correction type-sensitive normalization of english tweets using efficiently indexed n-gram statistics. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 120–125, Beijing, China. Association for Computational Linguistics.

G. Burnage, R.H Baayen, R. Piepenbrock, and H. van Rijn. 1990. *CELEX: A Guide for Users.* Centre for Lexical Information.

Eleanor Clark and Kenji Araki. 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual english. *Procedia - Social and Behavioral Sciences*, 27:2 – 11. Computational Linguistics and Related Fields.

Yerai Doval Mosquera, Jesús Vilares, and Carlos Gómez-Rodríguez. 2015. Lysgroup: Adapting a spanish microtext normalization system to english. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 99–105, Beijing, China. Association for Computational Linguistics.

Mark Dredze, David A Broniatowski, and Karen M Hilyard. 2016. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34(30):3441–2.

G Gonzalez-Hernandez, A Sarker, K O 'Connor, and G Savova. 2017. Capturing the Patient's Perspective: a Review of Advances in Natural Language Processing of Health-Related Text. *Yearbook of medical informatics*, pages 214–217.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 421–432, Jeju Island, Korea. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2013. Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology*, 4(1):1–27.

Xiaolei Huang, Michael C. Smith, Michael Paul, Dmytro Ryzhkov, Sandra Quinn, David Broniatowski, and Mark Dredze. 2017. Examining Patterns of Influenza Vaccination in Social Media. In *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*.

Ning Jin. 2015. NCSU-SAS-ning: Candidate generation and feature engineering for supervised lexical normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 87–92, Beijing, China. Association for Computational Linguistics.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55:73–81.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Comput. Surv.*, 24:377–439.

Kenneth H Lai, Maxim Topaz, Foster R Goss, and Li Zhou. 2015. Automated misspelling detection and correction in clinical free-text records. *Journal of Biomedical Informatics*, 55:188–195.

Samuel Leeman-Munk, James Lester, and James Cox. 2015. NCSU_SAS_SAM: Deep encoding and reconstruction for normalization of noisy text. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 154–161, Beijing, China. Association for Computational Linguistics.

Wookhee Min and Bradford Mott. 2015. NCSU_SAS_WOOKHEE: A deep contextual long-short term memory model for text normalization. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 111–119, Beijing, China. Association for Computational Linguistics.

Danielle L. Mowery, Brett R. South, Lee Christensen, Jianwei Leng, Laura Maria Peltonen, Sanna Salanterä, Hanna Suominen, David Martinez, Sumithra Velupillai, Noémie Elhadad, Guergana Savova, Sameer Pradhan, and Wendy W. Chapman. 2016. Normalizing acronyms and abbreviations to aid patient understanding of clinical texts: ShARe/CLEF eHealth Challenge 2013, Task 2. *Journal of Biomedical Semantics*.

National Cancer Institute. NCI Dictionary of Cancer Terms.

National Library of Medicine (US). RxNorm.

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, Rachel Ginn, and Graciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the*

American Medical Informatics Association: JAMIA, 22(3):671–81.

Albert Park, Andrea L Hartzler, Jina Huh, David W Mcdonald, and Wanda Pratt. 2015. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. *J Med Internet Res*, 17(212).

Jon Patrick, Mojtaba Sabbagh, Suvir Jain, and Haifeng Zheng. 2010. Spelling correction in clinical notes with emphasis on first suggestion accuracy. In *2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 2–8.

Michael J Paul and Mark Dredze. 2009. A Model for Mining Public Health Topics from Twitter. Technical report, Johns Hopkins University.

Martin Reynaert. 2005. *Text-Induced Spelling Correction*. Ph.D. thesis, Tilburg University.

Abeed Sarker. 2017. A customizable pipeline for social media text normalization. *Social Network Analysis and Mining*, 7(1):45.

Abeed Sarker, Rachel Ginn, Azadeh Nikfarjam, Karen O'Connor, Karen Smith, Swetha Jayaraman, Tejaswi Upadhaya, and Graciela Gonzalez. 2015. Utilizing social media data for pharmacovigilance: A review. *Journal of Biomedical Informatics*, 54:202–212.

Abeed Sarker and Graciela Gonzalez. 2017a. A corpus for mining drug-related knowledge from Twitter chatter: Language models and their utilities. *Data in Brief*, 10:122–131.

Abeed Sarker and Graciela Gonzalez. 2017b. Hlp@upenn at semeval-2017 task 4a: A simple, self-optimizing text classification system combining dense and sparse vectors. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 640–643, Vancouver, Canada. Association for Computational Linguistics.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety*, 39(3):231–240.

Dmitry Supranovich and Viachaslau Patsepnia. 2015. Ihs_rd: Lexical normalization for english tweets. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 78–81, Beijing, China. Association for Computational Linguistics.

Suzan Verberne. 2002. Context-sensitive spell checking based on word trigram probabilities. Master's thesis, Radboud University.

Nicole Walasek. 2016. Medical Entity Extraction on Dutch forum data in the absence of labeled training data. Master's thesis, Radboud University.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third social media mining for health (smm4h) shared tasks at emnlp 2018. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 13–16, Brussels, Belgium. Association for Computational Linguistics.

Q Zeng and T Tse. 2006. Exploring and developing consuming health vocabulary. *J Am Med Inform Assoc*, 13(1):24–29.

Xiaofang Zhou, An Zheng, Jiaheng Yin, Rudan Chen, Xianyang Zhao, Wei Xu, Wenqing Cheng, Tian Xia, and Simon Lin. 2015. Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care. *JMIR Med Inform 2015*, 3(3):e27.