Sentence Classification with Imbalanced Data for Health Applications

Farhana Ferdousi Liza
School of Computer Science and Electronic Engineering
University of Essex, Colchester, UK
farhana.ferdousi.liza@essex.ac.uk

Abstract

Identifying and extracting reports of medications, their abuse or adverse effects from social media is a challenging task. In social media, relevant reports are very infrequent, causes imbalanced class distribution for machine learning algorithms. Learning algorithms typically designed to optimize the overall accuracy without considering the relative distribution of each class. Thus, imbalanced class distribution is problematic as learning algorithms have low predictive accuracy for the infrequent class. Moreover, social media represents natural linguistic variation in creative language expressions. In this paper, we have used a combination of data balancing and neural language representation techniques to address the challenges. Specifically, we participated the shared tasks 1, 2 (all languages), 4, and 3 (only the span detection, no normalization was attempted) in Social Media Mining for Health applications (SMM4H) 2020 (Klein et al., 2020). The results show that with the proposed methodology recall scores are better than the precision scores for the shared tasks. The recall score is also better compared to the mean score of the total submissions. However, the F1-score is worse than the mean score except for task 2 (French).

1 Introduction

Advances in representation learning that attempts to automatically learn features for natural language processing (Young et al., 2018) present the possibility of utilizing social media (i.e. Twitter) data source for public health applications such as health monitoring and surveillance. Several ethical, legal and methodological challenges need to be addressed that are unique to Twitter data source (Ahmed et al., 2017). The ongoing shared tasks in Social Media Mining for Health applications (SMM4H) define evolving challenges specific to the Twitter data source for health domain (Weissenbacher et al., 2019). To address the methodological challenges, in recent years, several techniques have been proposed based on the SMM4H shared tasks (Sarker et al., 2018; Weissenbacher et al., 2019).

Epidemiologists intend to detect mentions of health issues related to medications early on Twitter. The adverse effect of medications is one of the leading causes of post-therapeutic deaths (Saha et al., 2018). One of the challenges of detecting real reports on medications, their abuse or adverse effects is to distinguish the relevant true reports from other general statements, news, and institutional advice. Reports on health issues (i.e. abuse or adverse effects of a medication) in social media are rare instances (i.e. very small percentage contain relevant information) (Weiss, 2004; Batista et al., 2004). This causes a major machine learning methodological challenges called imbalanced learning problem (He and Garcia, 2009; Ling and Sheng, 2010). The learning problem happens in the presence of underrepresented data and severely skewed class distribution, the situation where one of the class categories comprises a significantly larger proportion of the dataset than the other classes. Imbalanced class distribution is a practical issue in most real-world datasets (e.g. fraud detection, disease detection) and complicates learning when the identification of the minority class is of specific importance. This is a general problem for health domain with medical data utilizing machine learning methods (Rahman and Davis, 2013).

Proceedings of the 5th Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task, pages 138–145
Barcelona, Spain (Online), December 12, 2020.

When dealing with imbalanced data, model evaluation is different than typical class-balanced loss based machine learning framework (Cui et al., 2019; Marchand and Strawderman, 2020). Models that trained by minimizing errors on imbalanced datasets, tend to frequently predict the majority class; achieving high overall accuracy in such cases can be misleading. As class imbalance is a widespread issue, multiple techniques have been developed that help alleviate the issue (Buda et al., 2018; Haixiang et al., 2017), by either adjusting the model (e.g. changing the performance metric) or changing the data distribution (e.g. oversampling the minority class or undersampling the majority class). In this paper, we have explored novel strategies to handle extreme imbalanced class distribution.

Another challenge is representing informal Twitter text efficiently for health monitoring. Tweets contain misspelled unnormalized health concepts, expressed in a noisy, ungrammatical, multilingual, and ambiguous way. Moreover, creative and colloquial language expressions are prevalent in Twitter text. In this paper, advanced preprocessing and feature learning techniques were utilized to efficiently capture syntactic and semantic regularities of the substantially informal Twitter data. Overall, several techniques were explored to address challenges in Twitter data source for health monitoring and surveillance. The paper is organized as follows: section 2 describes data and problem statement, section 3 and 4 include the methodology used to model the datasets, and section 5 contains a discussion of the obtained results to understand the challenges specifically related to imbalanced dataset.

2 Problem Statement and Data Description

In this paper, six datasets from the Social Media Mining for Health application (SMM4H) 2020 Shared tasks were used for health monitoring and surveillance challenges. All the shared task datasets are labeled and can be modeled in one of three popular types of supervised classification problems. Binary (B) classification is the task of classifying the elements of a given set into two groups whereas multi-class (M) classification generalizes into more than two groups. Span detection (D) and normalization (N) task is defined as D+N which can be modeled as multi-class classification task followed by named-entity recognition modeling (Nadeau and Sekine, 2007; Lample et al., 2016).

Task ID: description	Type	IR (%)	# (*)	#Train	#Valid.	#Test
T1: Med. Mention	B	0.26	181 (1) 69,091 (0)	55419	13853	29687
T2 (En): AE in English	B	9.25	2,374 (1) 23,298(0)	20544	5134	4759
T2 (Fr): AE in French	B	1.61	39(1) 2,387(0)	1941	485	607
T2 (Ru): AE in Russian	B	8.75	666 (1) 6,946(0)	6090	1522	1903
T3: AE	D+N	51.20	1,212 (1) 1,155(0)	2246	560	976
T4: Med. Abuse	M	15.99	1685('a') 5488('m') 2940('c') 424('u')	10537	2635	3271

Table 1: Brief data description: rows correspond to the datasets and columns correspond to various attributes related to the datasets. The first column describes the task with an identifier. The second column denotes a classification type for the task (i.e. binary, multi-class and span-detection). The third column describes the imbalance ratio (IR), defined as the ratio of the minority class examples to the total number of examples, in percentage. The fourth column (# (*)) has the class/label distribution over the training dataset where #(*) denotes the number of Tweets labeled with *. Each dataset is comprised of three splits: training, validation and testing set. The fifth (i.e. #Train), sixth (i.e. #Valid.) and seventh (i.e. #Test) column correspond to the number of Tweets available for training, validation and testing set.

Table 1 shows a brief description of the datasets from the SMM4H 2020 shared tasks (Klein et al., 2020). Task 1 requires distinguishing between two classes of Tweets and modeled as a binary classification task. Positive Tweets that mention a medication or dietary supplement are labeled as "1" and

negative Tweets that do not mention are labeled as "0". The dataset consists of Tweets posted by 112 women during pregnancy, with approximately 0.26% (see column IR%) of the training Tweets mentioning a medication or dietary supplement. The data set represents an extremely[1] imbalanced class distribution.

Task 2 is another binary classification task that involves classifying Tweets based on the mentions of adverse effect (AE) of a medication. This task includes distinct sets of Tweets posted in three languages: English (En), French (Fr) and Russian (Ru). In Tab. 1, rows corresponding to task id T2 (En), T2 (Fr) and T2 (Ru) refer to sub-tasks in three languages. Tweets that mention an adverse effect of medications are labeled as "1" and those that do not have mention are labeled as "0". The tasks require taking into account subtle linguistic and semantic variations between AEs and indications (i.e. the reason for using the medication). The T2 (Fr), T2 (En) and T2 (Ru) datasets represent moderately imbalanced class distribution with IR of 1.61%, 9.25% and 8.75% respectively.

Task 3 involves detecting the span of Tweet containing an adverse effect (AE) of a medication and then mapping the extracted AE to a standard concept identifier (ID) in the MedDRA vocabulary (preferred terms). The training data includes Tweets that report or indicate an AE that are labeled as "1" and those that do not mention are labeled as "0". The detection task thus requires a model to distinguish between AEs and indications. The class distribution of the dataset is balanced (IR=51.20%) for the detection task. The normalization task involves classification to multiple classes where each class can be defined as a standard MedDRA ID. We have not attempted the normalization part of the task and left for future work.

Task 4 requires distinguishing between more than two classes of Tweets and can be modeled as a multi-class classification problem. The task involves distinguishing among Tweets that mention at least one prescription opioid, benzodiazepine, atypical anti-psychotic, central nervous system stimulant or GABA analogue. Tweets that report potential abuse/misuse are labeled as "A" from those that report non-abuse/-misuse consumption which are labeled as "C", merely mention of the medication are labeled as "M", and unrelated are labeled as "U". The task has moderately imbalanced class distribution with IR of 15.99%.

3 Methodology for Classification Tasks

We have developed systems for shared tasks 1, 2 (all languages), 4, and 3 (only the span detection, no normalization was attempted). The main focus was towards the imbalanced classification tasks. In this section, we will describe the methods used for five classification tasks and in the next section, we will describe the method developed for the span-detection part of task 3. The models were trained on the training dataset and evaluated on validation dataset; test data was not available when the methods were being developed. Each solution of the binary and multi-class classification task can be subdivided into four common steps and these steps will be elaborated in the following subsections.

3.1 Preprocessing

Tweet language processing is challenging compared to the standard text found in the news, journals and books (Balahur, 2013). For example, Twitter users use an informal language (Tan et al., 2015) that uses special expressions, such as "lol", "omg", emoticons (Derks et al., 2008), and emphasize or exaggerate the underlying meaning of the root word by using stretched words like 'heellllp' or 'heyyyyy'. Using the stretched word is common in spoken language but Tweets can have them in written format. The traditional mainstream natural language processing tools were not designed to include the syntactic regularities of informal language (Kong et al., 2014). Additionally, Twitter language has specific metadata, such as "RT" defines Tweets that reposted by other users, the markup of topics using the hashtag ("#") sign and defining other Twitter users by "@" sign. In this paper, we have considered the Tweet specific characteristics when preprocessing the Tweets so that the language can be normalized — converting them into a more standard form of language — efficiently. Normalization of Twitter posts enable us to apply standard natural language processing (NLP) techniques more effectively.

[1] https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

A basic normalization (Kaufmann and Kalita, 2010; Declerck and Lendvai, 2015; Beckley, 2015) was employed as part of advanced preprocessing step for the datasets. The preprocessing stage includes the following steps: separating hyperlinks from the adjacent text, normalize twitter-specific tokens, extracting text from '*' (e.g. *good* $->$ good), replacing & symbol, lower-casing the text, normalize multiple occurrences of vowels and consonants, normalize emojis and numbers, spiting 'number' and 'emoji' when adjacent to text, removing non-alphanumeric characters, removing very long words $>= 15$ and short words < 2 to reduce sparsity, removing multiple sequential occurrences of the same token.

3.2 Under-sampling

As one of the main focus of this paper is to model imbalanced dataset, we have utilized random resampling (Napierała et al., 2010; Johnson and Khoshgoftaar, 2019) to balance the class distribution. After Tweet preprocessing, we have used random under-sampling of the majority class. Oversampling is another approach to balance the training dataset, however, often oversampling minority class instances does not fully balance the training data whereas under-sampling of majority class is found to be better at balancing the training data (Jamil, 2017). We have left experiments involving oversampling for future works.

For extremely imbalanced class distribution in task 1, the under-sampling was done in two steps. First, we have utilized an informed under-sampling technique based on a pre-trained named-entity recognizer (NER) (Andriy Mulyar and McInnes, 2018) trained on clinical notes[2]. The NER model was used to extract and remove the Tweets that have medical named entities represented by informal expressions (e.g. 'lol').

After that, random under-sampling of the majority class was done on the remaining Tweets. The balanced dataset size is twice the number of the positive label in the dataset. For example, task 1 has a total of 146 positive examples (i.e. minority class) and after applying the random under-sampling, the negative examples (i.e. majority class) reduced down to 146. The resulting balanced training dataset size is 292 (i.e. 146 + 146). This part of random down-sampling was applied for all the datasets with IR $< 50\%$. A python toolbox, imbalanced-learn (Lemaître et al., 2017), was used to re-balance the class distribution.

3.3 Sentence Embedding

Sentence embedding methods attempt to encode a variable-length input sentence into a fixed-length vector. While preprocessing and normalization helps in better syntactic representation (Kaufmann and Kalita, 2010; Kong et al., 2014), sentence embeddings were used to improve the semantic representation. Moreover, sentence embeddings have been used in sentence classification to address the class imbalance problem (Madabushi et al., 2019). In recent years, several sentence embedding methods, exploration of semantic properties of resulted embeddings and impact of embedding on the downstream application have been proposed (Le and Mikolov, 2014; Kiros et al., 2015; Pagliardini et al., 2018; Schwenk and Douze, 2017; Arora et al., 2019; Zhu et al., 2018). Among them Sent2Vec (Pagliardini et al., 2018) demonstrated the robustness of generated general-purpose sentence embeddings when transferred to a wide range of prediction benchmarks. Sent2Vec is an unsupervised sentence embedding technique allowing composing sentence embeddings using word vectors along with n-gram embeddings.

There are three distinct language-specific categorizations of the Tweets in the shared task datasets. Among the participated tasks, four are based on English Tweets whereas two tasks are based on French and Russian Tweets. For representing the four English language Tweets, we have used 700-dimensional pre-trained[3] Sent2Vec model trained with English Tweets incorporating bi-gram embedding. For non-English Tweets, we have used 1024-dimensional pre-trained Language-Agnostic Sentence Representations (LASER) embeddings[4] (Schwenk and Douze, 2017). The pre-trained LASER model was based on 93 languages and does not need a specification of the input language. The sentence encoder also supports code-switching, i.e. the same sentences can contain words in several different languages. Overall,

[2]https://github.com/NLPatVCU/medaCy_model_clinical_notes

[3]https://github.com/epfml/sent2vec

[4]https://github.com/facebookresearch/LASER

Table 2: Evaluation Score on Validation, Test Datasets and Mean Score Based on all Submissions

	Validation			**Test**			**Mean (all submissions)**		
Task	F1	P	R	F1	P	R	F1	P	R
1	0.08	0.04	**0.97**	0.05	0.03	**0.90**	0.66	0.70	**0.69**
2 (EN)	0.39	0.25	**0.84**	0.32	0.19	**0.87**	0.46	0.42	**0.59**
2 (FR)	0.08	0.04	**0.75**	0.07	0.04	**0.60**	0.07	-	-
2 (RU)	0.35	0.22	**0.86**	0.35	0.22	**0.89**	0.43	0.36	**0.58**
3 (D)	-	-	-	0.159	**0.178**	0.143	0.564	**0.607**	0.557
4	0.45	0.36	**0.62**	0.46	0.35	**0.68**	0.49	-	-

Sent2Vec model was used to embed the English Tweets, whereas LASER was used for multilanguage tasks.

3.4 Classification Model Selection

We mainly focus on traditional systems to classify the Tweets. In recent years, deep learning models have shown superior performance in classification tasks (Weissenbacher et al., 2019). However, neural models require longer training time (Livni et al., 2014) and due to time constraint, we have utilized traditional baselines from related works (Weissenbacher et al., 2019). The multi-class classification is modeled as an one-vs-one scheme. We have applied Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with radial basis function kernel and tree-based ensemble models, such as, Extra Trees Classifier (Geurts et al., 2006), Random Forest Classifier (Breiman, 2001) for data modeling. The features for models were learned by sentence embedding models as described in Sec. 3.3 and the final model was chosen based on the 10 split k-fold cross-validation on the down-sampled training dataset. For most of the tasks, the SVM model provided better result based on 10-fold cross-validation with mean F1-score evaluation metric, except for T2 (Fr) task where Extra Tree Classifier gave best mean F-1 score. Based on the experiments, for evaluation of the shared tasks, Extra Tree Classifier was used for T2 (Fr), and SVM models were used for all other participated tasks.

4 Methodology for Span Detection Task

Task 3 can be divided into two sub-tasks: span detection and concept normalization. We have worked on the detection part of the task and the normalization part is left for the future work. The dictionary-based simple traditional approach has been used in information detection task (Egorov et al., 2004). The approach utilizes a carefully constructed dictionary to identify and tag the related entities from the Tweets. For task 3, we have used a dictionary-based search approach to detect the span of adverse effects. A dictionary was generated from the AMIA shared task[5] with 6,649 annotated instances[6] of adverse effects.

5 Results and Analysis

The official performance evaluation metrics for the tasks are precision (P), recall (R), and F1-score (F1) computed on the positive class (i.e. minority class). Table 2 reports the performance scores of the tasks described in Tab. 1. The rows of the table correspond to the tasks and the columns correspond to the evaluation scores on validation and test datasets. The mean test scores of all the submissions are also reported in the table. The analysis of the results is based on the validation dataset.

Using the methodology described in this paper, from the table we can observe that the recall score is higher compared to the precision score in most cases. As positive labels refer to the minority class, there are a larger number of negative examples that could become false positives (FP). Conversely, there are fewer positive examples that could become false negatives (FN). The disproportion in class distribution

[5]https://healthlanguageprocessing.org/sharedtask2/

[6]https://github.com/tttr222/uknlp_adr_mention_norm/tree/master/data_train

can result in the disproportion of the FPs and FNs which could result in high recall and low precision score. As F1-score is a weighted combination of these two matrices, a low precision score can cause an overall low F1-score. Although random under-sampling was used to balance the training dataset, it does not improve the precision score on the validation and test dataset since the validation and test dataset are still imbalanced that reflect the real class distribution.

6 Conclusion

This paper reports the preliminary exploration of Twitter-based imbalanced data modeling techniques for health applications. We have explored techniques for efficient syntactic and semantic representation of Tweets. As the focus was on imbalanced class distribution, random under-sampling with novel noise reduction technique was utilized to balance the dataset. The traditional margin-based and ensemble tree-based classifiers were used to classify the Tweets.

The directions for future works involve deeper exploration and extensive methodological improvements to enhance the learning performance. For example, the statistical learning theorems establish the number of examples needed to estimate the accuracy of a classifier as a function of its complexity (VC-dimension). However, the class imbalance does not enter these formulas anywhere (Juba and Le, 2019). Overall, a detailed exploration of classification models and evaluation metrics can be done to enhance model performance for skewed data distribution.

Acknowledgements

The author would like to acknowledge the support of the Business and Local Government Data Research Centre (ES/S007156/1) funded by the Economic and Social Research Council (ESRC) for undertaking this work.

References

Wasim Ahmed, Peter A Bath, and Gianluca Demartini. 2017. Using twitter as a data source: An overview of ethical, legal, and methodological challenges. In *The Ethics of Online Research*. Emerald Publishing Limited.

Natassja Lewinski Andriy Mulyar and Bridget McInnes. 2018. Tac srie 2018: Extracting systematic review information with medacy. *National Institute of Standards and Technology (NIST) 2018 Systematic Review Information Extraction (SRIE) ¿ Text Analysis Conference*, nov.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2019. A simple but tough-to-beat baseline for sentence embeddings. In *5th International Conference on Learning Representations, ICLR 2017*.

Alexandra Balahur. 2013. Sentiment analysis in social media texts. In *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 120–128.

Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. 2004. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29.

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

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277.

Thierry Declerck and Piroska Lendvai. 2015. Processing and normalizing hashtags. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 104–109, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2008. Emoticons and online message interpretation. *Social Science Computer Review*, 26(3):379–388.

Sergei Egorov, Anton Yuryev, and Nikolai Daraselia. 2004. A simple and practical dictionary-based approach for identification of proteins in medline abstracts. *Journal of the American Medical Informatics Association*, 11(3):174–178.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely randomized trees. *Machine learning*, 63(1):3–42.

Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

Haibo He and Edwardo A Garcia. 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284.

Zunaira Jamil. 2017. *Monitoring tweets for depression to detect at-risk users*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27.

Brendan Juba and Hai S Le. 2019. Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4039–4048.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*, volume 16.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

Ari Z. Klein, Ilseyar Alimova, Ivan Flores, Arjun Magge, Zulfat Miftahutdinov, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the fifth social media mining for health applications (# smm4h) shared tasks at coling 2020. *In Proceedings of the Fifth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task.*

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1188–1196, Bejing, China, 22–24 Jun. PMLR.

Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5.

Charles X. Ling and Victor S. Sheng, 2010. *Class Imbalance Problem*, pages 171–171. Springer US, Boston, MA.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. 2014. On the computational efficiency of training neural networks. In *Advances in neural information processing systems*, pages 855–863.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive bert for generalisable sentence classification with imbalanced data. *EMNLP-IJCNLP 2019*, page 125.

Éric Marchand and William E Strawderman. 2020. On shrinkage estimation for balanced loss functions. *Journal of Multivariate Analysis*, 175:104558.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, January. Publisher: John Benjamins Publishing Company.

Krystyna Napierała, Jerzy Stefanowski, and Szymon Wilk. 2010. Learning from imbalanced data in presence of noisy and borderline examples. In *International Conference on Rough Sets and Current Trends in Computing*, pages 158–167. Springer.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

M Mostafizur Rahman and Darryl N Davis. 2013. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224.

Rupsa Saha, Abir Naskar, Tirthankar Dasgupta, and Lipika Dey. 2018. Leveraging web based evidence gathering for drug information identification from tweets. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 67–69.

Abeed Sarker, Maksim Belousov, Jasper Friedrichs, Kai Hakala, Svetlana Kiritchenko, Farrokh Mehryary, Sifei Han, Tung Tran, Anthony Rios, Ramakanth Kavuluru, Berry de Bruijn, Filip Ginter, Debanjan Mahata, Saif M Mohammad, Goran Nenadic, and Graciela Gonzalez-Hernandez. 2018. Data and systems for medication-related text classification and concept normalization from Twitter: insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *Journal of the American Medical Informatics Association*, 25(10):1274–1283, 10.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167.

Luchen Tan, Haotian Zhang, Charles Clarke, and Mark Smucker. 2015. Lexical comparison between wikipedia and twitter corpora by using word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 657–661.

Gary M Weiss. 2004. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen OConnor, Michael Paul, and Graciela Gonzalez. 2019. Overview of the fourth social media mining for health (smm4h) shared tasks at acl 2019. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 21–30.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2018. Recent trends in deep learning based natural language processing. *IEEE Computational intelligenCe magazine*, 13(3):55–75.

Xunjie Zhu, Tingfeng Li, and Gerard De Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637.