Creating and Evaluating Resources for Sentiment Analysis in the Low-resource Language: Sindhi

Wazir Ali[1] **Naveed Ali**[1] **Yong Dai**[1] **Jay Kumar**[1]

Saifullah Tumrani[1] **Zenglin Xu**[1,2]

[1]University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Harbin Institute of Technology, Shenzhen, Nanshan 510085, China
{aliwazirjam,zenglin}@gmail.com

Abstract

In this paper, we develop Sindhi subjective lexicon using a merger of existing English resources: NRC lexicon, list of opinion words, SentiWordNet, Sindhi-English bilingual dictionary, and collection of Sindhi modifiers. The positive or negative sentiment score is assigned to each Sindhi opinion word. Afterwards, we determine the coverage of the proposed lexicon with subjectivity analysis. Moreover, we crawl multi-domain tweet corpus of news, sports, and finance. The crawled corpus is annotated by experienced annotators using the Doccano text annotation tool. The sentiment annotated corpus is evaluated by employing support vector machine (SVM), recurrent neural network (RNN) variants, and convolutional neural network (CNN).

1 Introduction

The exponential growth in the online professional and user-generated textual data (Akhtar et al., 2016), including blog posts, news headlines, product, and book reviews, led to the growth of the sentiment analysis task. The required essential resources for the classification of such opinionated text are the polarity assigned sentiment lexicon (Asghar et al., 2019) and sentiment annotated corpora (Ekbal et al., 2020). Sophisticated research efforts have been employed for English sentiment analysis (Joshi et al., 2017; Hussein, 2018). In the result, a number of resources are available including opinion words (Hu and Liu, 2004), subjective lexicon (Wilson et al., 2005), SentiWordNet (SWN) (Esuli and Sebastiani, 2006; Baccianella et al., 2010), NRC lexicon (Mohammad and Turney, 2010). The sentiment annotated corpora including financial news (FN) (Takala et al., 2014), sports tweets (ST) (Yu and Wang, 2015), tweet dataset (Thelwall et al., 2012), and more recently a multi-domain corpora (Ekbal et al., 2020). Among these resources the SWN has been widely used for the construction of sentiment lexicon for low-resource languages including Urdu (Asghar et al., 2019), Turkish (Dehkharghani et al., 2016), and Hindi (Bakliwal et al., 2012).

Sindhi is an Indo-Aryan language, spoken by more than 75 million (Motlani, 2016) people. Presently, it is being written in two main scripts of Persian-Arabic and Devanagari (Jamro, 2017). However, Persian-Arabic is a popular and standard script (Ali et al., 2020). It is widely used in online communication, mainly in the Sindh province of Pakistan and some regions of India (Ali et al., 2019). The generated content on social media contains rich information about the interests of individuals. Thus, the modeling of such information is essential to analyze where people's opinions are conveyed. The low-resource Sindhi language lacks the primary resources for content analysis, such as polarity assigned lexicon and sentiment annotated corpora.

In this paper, we create Sindhi subjective lexicon using existing English resources. Moreover, due to the scarcity of sentiment annotated corpora, we crawl and annotate news headline (NH) tweets, sports tweets (ST), and FN tweets, respectively. Three native annotators performed the annotation using Doccano (Nakayama et al., 2018) text annotation tool with 79.3% inter-annotator agreement. To the best of our knowledge, both datasets[1] are the first benchmark for Sindhi sentiment analysis (SSA). Furthermore, we develop strong baselines of SVM (Cortes and Vapnik, 1995), CNN (Dos Santos and Gatti, 2014) and RNN variants of long-short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), bidirectional long-short-term memory (BiLSTM) (Schuster and Paliwal, 1997) for the evaluation purpose.

[1] The resources can be found at https://github.com/AliWazir/SdSenti-lexicon

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 188–194
April 19, 2021. ©2021 Association for Computational Linguistics

2 Related Work

The development of polarity assigned sentiment lexicon have largely been investigated for the rich-resource English language, such as Bing Liu's lexicon (Hu and Liu, 2004), SentiWordNet (SWN) (Esuli and Sebastiani, 2006), SWN 3.0 (Baccianella et al., 2010), and NRC lexicon (Mohammad and Turney, 2010). These resources have been widely used to create a sentiment lexicon for low-resource languages, mainly by translating the terms into the target languages. The human annotators assigned polarity score to create the sentiment lexicon for South Asian languages such as Hindi (Bakliwal et al., 2012), Bengali, Telugu (Das and Bandyopadhyay, 2010), Tamil (Kannan et al., 2016), Persian (Amiri et al., 2015), Urdu (Asghar et al., 2019), Panjabi (Kaur and Gupta, 2014), and Sinhala (Medagoda et al., 2015). Bakliwal et al. (2012) proposed a graph-based WordNet-based approach to develop subjective lexicon by using synonym and antonym relations. Das and Bandyopadhyay (2010) opted multiple methods such as, dictionary-based, corpus-based or generative approach, and WordNet-based for the construction of sentiment lexicon for Indian languages. Various resources and tools including Bing Liu's lexicon, SWN, subjectivity lexicon (Wilson et al., 2005), AFINN-111 lexicon (Nielsen, 2011), and Google translate (Amiri et al., 2015) are utilized to develop sentiment lexicon for Persian language. A word-level translation scheme (Asghar et al., 2019) is proposed to construct Urdu lexicon using English resources including Bing Liu's lexicon, SWN 3.0, and English to Urdu bilingual dictionary. Medagoda et al. (2015) proposed sentiment lexicon for low-resource Sinhala Language by using SWN 3.0 using word-level translation scheme. As we mentioned earlier, Sindhi stands among the low-resource languages because lack the subjective lexicon and sentiment annotated corpus except a recently Ali and Wagan 6842 part-of-speech tagged lexicon. Hence, their lexicon lack a sentiment intensity score. Thus, we propose Sindhi subjective lexicons using a merger of existing English resources. Moreover, we also crawl multi-domain NH, FT, and ST tweets and annotate them using Doccano (Nakayama et al., 2018) text annotation tool.

Many sentiment-annotated corpora have also been created for English in multiple domains, such as news, sports, finance, and products. Shamma et al. (2009) created tweet dataset by crawling U.S. presidential debate in multiple sentiment classes. Blitzer et al. (2007) proposed sentiment dataset on product reviews, electronics, and kitchen appliances obtained from Amazon.com. (Thelwall et al., 2012) manually annotated tweet dataset with $+ve$ and $-ve$ sentiments. The annotation of financial blogs and news domain (O'Hare et al., 2009; Malo et al., 2013; Takala et al., 2014) have also been investigated at a large scale. Yu and Wang (2015) proposed a dataset by crawling sports tweets from Twitter using search API. More recently, Ekbal et al. (2020) proposed multi-domain tweet corpora, annotated with three sentiment classes. Review shows that Sindhi lacks the subjective lexicon as well as sentiment annotated corpora for its supervised sentiment analysis that we consider.

3 Development of Subjective Lexicon

We construct Sindhi subjective lexicon by depicting the sentiment polarity score of all English opinion words using bilingual English to Sindhi dictionary[2]. The construction steps are described as follows:

3.1 Used Resources

To create the Sindhi subjective lexicon, we merge the list of Bing Liu's opinion words and the NRC lexicon. Afterwards, sentiment polarity is assigned using SWN 3.0 and translated to Sindhi using a bilingual dictionary.

- **NRC lexicon** is the list of English opinion words associated with basic emotions of fear, anger, sadness, disgust, surprise, and joy, etc. The lexicon include 2,312 $+ve$ and 3,324 $-ve$ words.

- **Bing Liu's lexicon** are general purpose English sentiment lexicon consists of 2,036 $+ve$ 4,814 $-ve$ words.

- **SentiWordNet 3.0** contains 117,659 English WordNet synset. Each term is associated with a numerical opinion score ranging between $[0.0, 1]$ to indicate the sentiment strength into $+ve$, $-ve$, or neutral classes.

- **Sindhi modifiers** increase or decrease sentiment strength of opinion words. Thus, we collect 173 Sindhi modifiers and assigned polarity using SWN 3.0 as well as human judgment. We manually assign the score to modifiers (see Table

[2] http://dic.sindhila.edu.pk/

5) in case of the unavailability of English translation of Sindhi modifiers in SWN 3.0 dataset.

– **English-Sindhi dictionary** is used to translate each English opinion word to the corresponding Sindhi word using comprehensive online English to Sindhi dictionary. If a bilingual dictionary returns more than one meaning of an opinion word, then the first or exact meaning is chosen by ignoring less common meanings.

3.2 Scoring Mechanism

We merge Bing Liu's, NRC lexicon and remove duplicates to develop a list of opinion lexicon. Each word from the list is looked up into SWN 3.0 to assign a polarity score. We choose the maximum polarity score of a retrieved word, such as we select -0.778 among all the synset of a word *heinous* in SWN (see Table 1). Afterwards, Sindhi translation is looked up in English to Sindhi dictionary. If a bilingual dictionary returns more than one meaning of a word, then the first or exact meaning (see Table 2) is chosen by ignoring less common or poetic meanings. An example of few constructed Sindhi subjective unigram and bigram terms is given in Table 3 and Table 4, respectively. Moreover, the sentiment score to Sindhi modifiers is assigned using SWN 3.0 and with decision making by assigning $+ve$ and $-ve$ polarity. Four native experienced annotators assigned polarity scores to the opinion lexicon and translated them into Sindhi. The overall inter-annotator agreement of 84.7% is achieved.

SenseID	Synsets	Pos	Neg	Neu
02514380	heinous#1	0.222	0.778	0.00

Table 1: An example of a word *heinous* (synset ID-02514380) in SWN 3.0 with two polarity scores.

Term	Synsets	Polarity score
Heinous	ڪِريل#3 مڪروه#2 سنگين #1	سنگين#1|-0.778

Table 2: An example of a translated English word *heinous* to its equivalent Sindhi word by choosing the first meaning.

Term	Sense ID	English Translation	Polarity Score
سٺو	01123148	Good	0.75
محفوظ	02550868	Save	0.50
بدلو	01153486	Revenge	-0.50
شرم	02547225	Shame	-0.625
پريشاني	02460502	True	0.50
تيزاب	1460752	Acid	-0.25
ٿڌ	01251128	Cold	-0.75

Table 3: List of few unigrams in the proposed Sindhi subjective lexicon. The *Sense ID* represents a WordNet (3.0) synset.

Terms	Sense ID	English Translation	Polarity Score
نا معلوم	00028672	Unacknowledged	-0.625
غير منظم	00641944	Unmannered	-0.625
نا اميد	01229020	Hopeless	-0.75
مقابلو ڪندڙ	00007990	Resistant	-0.5
ڦيريءَ بابت	02708232	Cyclic	0.5
انصاف ڪرڻ	05615373	Judiciousness	0.875
نا اهلي	05648953	Inefficiency	-0.50
تعريف جوڳو	02585545	Praiseworthy	0.625

Table 4: List of few bigrams in proposed Sindhi subjective lexicon. The *Sense ID* represents a WordNet (3.0) synset.

4 Development of Sentiment-annotated Corpus

Our main contributions include: a) The construction of polarity assigned Sindhi subjective lexicon using a merger of existing English resources. b) The acquisition of multi-domain NH, ST, FN tweet corpus and annotation for SSA using Doccano text annotation tool. c) The coverage of the proposed lexicon is determined with a subjectivity analysis test, and the sentiment annotated corpus is evaluated by employing SVM, LSTM, BiLSTM, and CNN models.

4.1 Data Acquisition

Due to the scarcity of corpus in multiple domains with gold annotations, we crawl the data from twitter using search API[3] and use web-scrapy[4,5] to collect the NH, ST, and FN headlines, tweets for sentiment annotation[6] text (see Table 7). The NH and ST tweets reflect the events, people's opinions, and their feelings about the events. The FN tweets contain people's opinions about inflation,

[3] https://twitter.com/search-advanced?f=live

[4] https://github.com/scrapy/scrapy

[5] List of few resources..
https://twitter.com/sindhtvnews247
https://twitter.com/awamiawazsindhi
https://www.pahenjiakhbar.com/categories/ trade
http://wichaar.com/news/202/ Finance & Business
http://wichaar.com/news/208/sports

[6] Corpus is crawled between December 11, 2019 to February 15, 2020

Term	Roman Transliteration	Polarity Score
گھٽ	Ghatt	-0.50
نا	Na	-0.50
بي	Bey	-0.375
اڻ	Annh	-0.25
تمام	Tamam	0.5
انتهائي	Intihayi	0.75

Table 5: List of few $+ve$, $-ve$ Sindhi modifiers.

Lexicon	Positive	Negative
Unigrams	3,986	7,562
Bigrams	179	269
Total	4,165	7,831

Table 6: Statistics of the proposed Sindhi subjective lexicon including modifiers.

the economy, capital expenditures, etc.

Dom	Tws	Sent	Pos	Neg	Neu
NH	2,096	3,534	1,134	1,141	1,259
ST	2,187	3,217	1,073	1,076	1,068
FN	1,754	2,853	953	952	948
Total	6,037	9,604	3,160	3,169	3,275

Table 7: Statistics of the preprocessed crawled corpus. The Dom, Tws, Sent, denote domain, tweets, and sentences. While Pos, Neg, Neu represent positive, negative, and neutral classes of the annotated sentences.

4.2 Data Preprocessing

We design a preprocessing pipeline for the filtration of unwanted data in the crawled tweets to get the desirable text for annotation, which consists of: a) Removal of unwanted punctuation marks from the start and end of the tweets. b) Filtration of noisy data such as special characters, non-Sindhi words, HTML tags, emails, and URLs. c) Normalization, removal of duplicates, multiple white spaces, and tweets that only contain user mentions. We also remove sentences containing more than 80 words and less than the length of 5 words.

4.3 Data Annotation

We use Doccano (Nakayama et al., 2018) text annotation tool for sentiment annotations of tweets into $+ve$, $-ve$, and neutral classes using crawled corpus (see Table 7). It is an open-source annotation tool for sequence labeling and sentiment analysis. The annotation is performed by three expert native annotators, keeping in view the sentiment ambiguities (Mohammad, 2016) in expressions such as success or failure, ridiculous expressions differing multiple entities, rhetorical questions, and requests are particularly challenging for the sentiment annotation. The overall inter annotation agreement (Cohen, 1960) of 79.3% shows the acceptable quality of the proposed dataset.

5 Evaluation

We determine the coverage of the proposed lexicon with subjectivity analysis (Asghar et al., 2019). The sentiment annotated corpus is evaluated by employing SVM, LSTM, BiLSTM, and CNN models.

5.1 Experimental Setup

The SVM, LSTM, BiLSTM, and CNN models are employed to evaluate the annotated dataset after combining all the domains. We filter stop words (Ali et al., 2019) and conduct the experiments, where the dataset was split into training, validation, and test sets. The results are reported in macro precision (P), recall (R), and F-value (F) for the average of the 10-fold runs along with accuracy (Kim, 2014) over the testing fold.

5.1.1 Representation Learning

A neural network requires word embedding (or sentence embedding) as an input to the network, i.e., a vector representation of each word or sentence. The tweets, books and news corpus (2174K tokens) (Ali et al., 2019) are converted into word representations before training neural models. The pretrained sub-word based representation learning has the ability to encode the structure of words (Bojanowski et al., 2017) at character-level by sharing the character n-gram representations across words. In that way, the representation for each word is made of the sum of those character n-grams. We obtain contextual representations by concatenating the sentence representations obtained through both $\overrightarrow{h}$ and $\overleftarrow{h}$ of BiLSTM hidden layers and residual connection (Jiang et al., 2019).

5.1.2 Support Vector Machine

We Employ SVM (Cortes and Vapnik, 1995) as an initial baseline for opinion extraction in each domain. The input features include N-gram tokens $(N = 1, 2, 3)$, character N-grams $(N = 2, 3, 4, 5)$ and proposed lexicons to extract the features.

5.1.3 Deep Neural Models

We employ LSTM, BiLSTM, and CNN models for the evaluation of our proposed dataset. The LSTM, BiLSTM networks can learn long-term dependencies. They contain input, forget, and output gates, which determine how much information should be lost and how much information should be added to memory. The BiLSTM network has the ability to encode past (left) and future (right) contexts in two separate forward and backward hidden states. Then both hidden states are concatenated for the final output. Moreover, the CNN consists of a representation layer, two convolutional layers, a pooling layer, and a fully connected layer.

5.1.4 Training Methodology

For the training of neural models for sentence type classification, all the sentences of each domain are used. The LSTM contains 300 hidden layers, and BiLSTM has 300 forward and 300 backward hidden layers, the concatenation of both resulted in 600 layers. A dense layer follows each hidden unit. We project input features by utilizing the dense layer. We employ 0.25% dropout in the fully connected LSTM, BiLSTM layers and 0.50% for CNN (Srivastava et al., 2014) and Adam optimizer (Kingma and Ba, 2015) with learning rate of 0.001%. All the neural models are implemented using TenserFlow (Abadi et al., 2016) deep learning framework on GTX 1080-TITAN GPU.

6 Results and Analysis

To assess the coverage of the proposed lexicon, the subjectivity analysis experiment is conducted to classify the sentences as subjective or objective. The sentence is classified as subjective if it contains one or more subjective word(s), otherwise classified as an objective in the absence of subjective word(s). The classification results of each domain are depicted in Table 8. Afterwards, all the domains are combined for sentence-level classification using supervised classifiers of SVM, LSTM, BiLSTM, and CNN, respectively. The overall performance of the SVM and neural models is presented in Table 9. The SVM is the weakest baseline classifier. It yields an accuracy of 67.86% with 68.00% precision, 69.00% recall, and 68.00% F1-value. The LSTM network shows better results than SVM by outputting 81.42% precision, 82.59% recall, 81.76% F-value, and 79.83% accuracy. The BiLSTM yields the best F-value of 83.11% and accuracy of 82.37%, respectively. The performance of the CNN network is very close to BiLSTM with precision 83.26%, recall 82.67%, F1-value 82.54%, and accuracy 81.68%.

Domain	P(%)	R(%)	F(%)
NH	75.28	74.65	73.89
ST	76.52	75.84	75.24
FN	75.33	75.69	74.61

Table 8: Results of subjectivity analysis.

Model	P(%)	R(%)	F(%)	A(%)
SVM	68.00	69.00	68.00	67.86
LSTM	81.42	82.59	81.76	79.83
BiLSTM	**83.70**	**84.37**	**83.11**	**82.37**
CNN	**83.26**	82.67	82.54	81.68

Table 9: The evaluation results based on supervised classifiers. The bold results reflect best performance.

The results demonstrate that the BiLSTM and CNN yield better results than SVM and LSTM. However, the BiLSTM network surpasses SVM, LSTM as well as CNN models.

7 Conclusion and Future Work

In this paper, we propose Sindhi subjective lexicon using various resources and sentiment annotated corpus, which serves as a benchmark for future expansions. The SVM and deep neural models are exploited for evaluation purposes. We achieve notable F-value and accuracy of 83.11%, 82.37% with the BiLSTM network. In the future, the proposed lexicon can be expanded using a corpus-based approach to capture language-specific words. Also, it can be used as a seed list, and the corpus can be tagged on the basis of the seed list. Moreover, the proposed lexicon consists of positive and negative classes, so a five-point scale can replace this classification in the future.

Acknowledgments

We would like to thank the anonymous reviewers for the helpful discussions and suggestions. This work was funded by the National Key R&D Program of China (No. 2018YFB1005100 & No. 2018YFB1005104).

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation*, pages 265–283.

Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in hindi: resource creation and evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2703–2709.

Mazhar Ali and Asim Imdad Wagan. Sentiment summerization and analysis of Sindhi text. *International Journal of Advanced Computer Science and Applications*.

Wazir Ali, Jay Kumar, Junyu Lu, and Zenglin Xu. 2019. Word embedding based new corpus for low-resourced language: Sindhi. *arXiv preprint arXiv:1911.12579*.

Wazir Ali, Jay Kumar, Zenglin Xu, Congjian Luo, Junyu Lu, Junming Shao, Rajesh Kumar, and Yazhou Ren. 2020. A subword guided neural word segmentation model for sindhi. *arXiv preprint arXiv:2012.15079*.

Fatemeh Amiri, Simon Scerri, and Mohammadhassan Khodashahi. 2015. Lexicon-based sentiment analysis for Persian text. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 9–16.

Muhammad Zubair Asghar, Anum Sattar, Aurangzeb Khan, Amjad Ali, Fazal Masud Kundi, and Shakeel Ahmad. 2019. Creating sentiment lexicon for sentiment analysis in Urdu: The case of a resource-poor language. *Expert Systems*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Akshat Bakliwal, Piyush Arora, and Vasudeva Varma. 2012. Hindi subjective lexicon: A lexical resource for Hindi polarity classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 1189–1196.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Amitava Das and Sivaji Bandyopadhyay. 2010. SentiWordNet for Indian languages. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 56–63.

Rahim Dehkharghani, Yucel Saygin, Berrin Yanikoglu, and Kemal Oflazer. 2016. SentiTurkNet: a Turkish polarity lexicon for sentiment analysis. *Language Resources and Evaluation*, 50(3):667–685.

Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *25th International Conference on Computational Linguistics*, pages 69–78.

Asif Ekbal, Pushpak Bhattacharyya, Shikha Srivastava, Alka Kumar, Tista Saha, et al. 2020. Multi-domain tweet corpora for sentiment analysis: Resource creation and evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5046–5054.

Andrea Esuli and Fabrizio Sebastiani. 2006. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 417–422.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Doaa Mohey El-Din Mohamed Hussein. 2018. A survey on sentiment analysis challenges. *Journal of King Saud University-Engineering Sciences*, 30(4):330–338.

Wazir Ali Jamro. 2017. Sindhi language processing: A survey. In *International Conference on Innovations in Electrical Engineering and Computational Technologies*, pages 1–8.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6281–6286.

Aditya Joshi, Pushpak Bhattacharyya, and Sagar Ahire. 2017. Sentiment resources: Lexicons and datasets. In *A Practical Guide to Sentiment Analysis*, pages 85–106. Springer.

Abishek Kannan, Gaurav Mohanty, and Radhika Mamidi. 2016. Towards building a SentiWordNet for Tamil. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 30–35.

Amandeep Kaur and Vishal Gupta. 2014. Proposed algorithm of sentiment analysis for Punjabi text. *Journal of Emerging Technologies in Web Intelligence*, 6(2):180–183.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Pekka Malo, Ankur Sinha, Pyry Takala, Oskar Ahlgren, and Iivari Lappalainen. 2013. Learning the roles of directional expressions and domain concepts in financial news analysis. In *IEEE International Conference on Data Mining Workshops*, pages 945–954.

Nishantha Medagoda, Subana Shanmuganathan, and Jacqueline Whalley. 2015. Sentiment lexicon construction using sentiwordnet 3.0. In *International Conference on Natural Computation*, pages 802–807.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 174–179.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Raveesh Motlani. 2016. Developing language technology tools and resources for a resource-poor language: Sindhi. In *Proceedings of the NAACL Student Research Workshop*, pages 51–58.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Workshop on'Making Sense of Microposts: Big things come in small packages*, pages 93–98.

Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F Smeaton. 2009. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 9–16.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

David A Shamma, Lyndon Kennedy, and Elizabeth F Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, pages 3–10.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Pyry Takala, Pekka Malo, Ankur Sinha, and Oskar Ahlgren. 2014. Gold-standard for topic-specific sentiment analysis of economic texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, volume 2014, pages 2152–2157.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005. Opinionfinder: A system for subjectivity analysis. In *HLT/EMNLP*, pages 34–35. The Association for Computational Linguistics.

Yang Yu and Xiao Wang. 2015. World cup 2014 in the twitter world: A big data analysis of sentiments in us sports fans tweets. *Computers in Human Behavior*, 48:392–400.