

# A system for the 2019 Sentiment, Emotion and Cognitive State Task of DARPA’s LORELEI project

Victor R Martinez\*, Anil Ramakrishna, Ming-Chang Chiu, Karan Singla and Shrikanth Narayanan†

Department of Computer Science,

University of Southern California

\*victorrm@usc.edu, †shri@sipi.usc.edu

**Abstract**—During the course of a Humanitarian Assistance-Disaster Relief (HADR) crisis, that can happen anywhere in the world, real-time information is often posted online by the people in need of help which, in turn, can be used by different stakeholders involved with management of the crisis. Automated processing of such posts can considerably improve the effectiveness of such efforts by, for example, understanding the aggregated emotion from affected populations in specific areas, which may help inform decision-makers on how to best allocate resources for an effective disaster response. However, these efforts may be severely limited by the availability of resources for a particular local language. The ongoing DARPA project *Low Resource Languages for Emergent Incidents* (LORELEI) aims to further language processing technologies for low resource languages in the context of such a humanitarian crisis. In this work, we describe our submission for the 2019 Sentiment, Emotion and Cognitive state (SEC) pilot task of the LORELEI project. We describe a collection of sentiment analysis systems included in our submission along with the features extracted. Our fielded systems obtained the best results in both English and Spanish language evaluations of the SEC pilot task.

## I. INTRODUCTION

The growing adoption of online technologies has created new opportunities for emergency information propagation [1]. During crises, affected populations post information about what they are experiencing, what they are witnessing, and relate what they hear from other sources [2]. This information contributes to the creation and dissemination of situational awareness [1], [3]–[5]. Crisis response agencies such as government departments or public health-care NGOs can make use of these channels to gain insight into the situation as it unfolds [3], [6]. Additionally, these organizations might also post time-sensitive crisis management information to help with resource allocation and provide status reports [7]. While many of these organizations recognize the value of the information found online—specially during the on-set of a crisis—they are in need of automatic tools that locate actionable and tactical information [1], [8].

Opinion mining and sentiment analysis techniques offer a viable way of addressing these needs, with complementary insights to what keyword searches or topic and event extraction might offer [9]. Studies have shown that sentiment analysis of social media during crises can be useful to support response coordination [10] or provide information about which audiences might be affected by emerging risk events [11]. For example, identifying tweets that convey a sense of “fear” might

support responders on assessing mental health effects among the affected population [12]. Given the critical and global nature of the HADR events, tools must process information quickly, from a variety of sources and languages, making it easily accessible to first responders and decision makers for damage assessment and to launch relief efforts accordingly [13], [14]. However, research efforts in these tasks are primarily focused on high resource languages such as English, even though such crises may happen anywhere in the world.

The LORELEI program provides a framework for developing and testing systems for real-time humanitarian crises response in the context of low-resource languages. The working scenario is as follows: a sudden state of danger requiring immediate action has been identified in a region which communicates in a low resource language. Under strict time constraints, participants are expected to build systems that can: translate documents (as necessary), find the named entities actively involved in the situation, and identify the underlying situation [15]. Situational information is encoded in the form of Situation Frames — data structures with fields identifying and characterizing the crisis type. The program’s objective is the rapid deployment of systems that can process text or speech audio from a variety of sources, including newscasts, news articles, blogs and social media posts, all in the local language, and populate these Situation Frames. While the task of identifying Situation Frames is similar to existing tasks in literature (e.g., slot filling), it is defined by the very limited availability of data [16]. This lack of data requires the use of simpler but more robust models and the utilization of transfer learning or data augmentation techniques.

The Sentiment, Emotion, and Cognitive State (SEC) evaluation task was a recent addition to the LORELEI program introduced in 2019, which aims to leverage sentiment information from the incoming documents. This in turn may be used in identifying severity of the crisis in different geographic locations for efficient distribution of the available resources. In this work, we describe our systems for targeted sentiment detection for the SEC task. Our systems are designed to identify authored expressions of sentiment and emotion towards a HADR crisis. To this end, our models are based on a combination of state-of-the-art sentiment classifiers and simple rule-based systems. We evaluate our systems as part of the NIST LoREHLT 2019 SEC pilot task.

## II. PREVIOUS WORK

Social media has received a lot of attention as a way to understand what people communicate during disasters [12], [17]. These communications typically center around collective sense-making [18], supportive actions [19], [20], and social sharing of emotions and empathetic concerns for affected individuals [21]. To organize and make sense of the sentiment information found in social media, particularly those messages sent during the disaster, several works propose the use of machine learning models (e.g., Support Vector Machines, Naive Bayes, and Neural Networks) trained on a multitude of linguistic features<sup>1</sup>. These features include bag of words, part-of-speech tags, n-grams, and word embeddings; as well as previously validated sentiment lexica such as Linguistic Inquiry and Word Count (LIWC) [23], AFINN [24], and SentiWordNet [25]. Most of the work is centered around identifying messages expressing sentiment towards a particular situation as a way to distinguish crisis-related posts from irrelevant information [26]. Either in a binary fashion (positive vs. negative) (e.g., [26]) or over fine-grained emotional classes<sup>2</sup> (e.g., [17]).

In contrast to social media posts, sentiment analysis of news articles and blogs has received less attention [27]. This can be attributed to a more challenging task stemming from the domain itself since, for example, journalists will often refrain from using clearly positive or negative vocabulary when writing news articles [28]. In spite of that, certain segments of this domain are still apt for sentiment analysis, such as column pieces [29] or political news [28], [30].

In the context of leveraging the information found online for HADR emergencies, approaches for languages other than English have been limited. Most of which are done by manually constructing resources for a particular language (e.g., in tweets [31]–[33] and in disaster-related news coverage [34]), or by applying cross-language text categorization to build language-specific models [32], [35].

In this work, we develop systems that identify positive and negative sentiments expressed in social media posts, news articles and blogs in the context of a humanitarian emergency. Our systems work for both English and Spanish by using an automatic machine translation (MT) system. We selected this approach on the basis that MT has been successfully leveraged for other low-resource languages (e.g., Uyghur) in the context for the SF prediction task [16], [36]. This also comes with the benefit of being easily extendable to other languages, bypassing the scalability issues that arise from the need to manually construct lexica resources.

## III. PROBLEM DEFINITION

This section describes the SEC task in the LORELEI program along with the dataset, evaluation conditions and metrics.

<sup>1</sup>For an in-depth review of these approaches, we refer the reader to [22]

<sup>2</sup>For example, anger, disgust, fear, happiness, sadness, and surprise

TABLE I  
EXAMPLE OF A SEGMENT CONTAINING A SITUATION FRAME WITH SENTIMENT RELATED ANNOTATIONS (IN BOLD).

Original Text	La crisis política que comenzó en abril pasado en Nicaragua, una situación inédita en la historia reciente del país, reporta al menos 79 muertos y 868 heridos, según cifras de la Comisión Interamericana de Derechos Humanos
Translation	The political crisis that began last April in Nicaragua, a situation unprecedented in the recent history of the country, reports at least 79 deaths and 868 wounded, according to figures from the Inter-American Commission on Human Rights.
Entities	Nicaragua (GPE), Comisión Interamericana de Derechos Humanos (ORG)
SF-Type	Medical Need
Location	Nicaragua
Status	Current, No known resolution, Non-urgent
<b>Sentiment</b>	Negative (-3)
<b>Emotion</b>	None
<b>Source</b>	Author
<b>Target</b>	Situation Frame

### A. The Sentiment, Emotion and Cognitive State (SEC) Task

Given a dataset of text documents and manually annotated situation frames, the task is to automatically detect sentiment polarity relevant to existing frames and identify the source and target for each sentiment instance. The source is defined as a person or a group of people expressing the sentiment, and can be either a PER/ORG/GPE (person, organization or geo political entity) construct in the frame, the author of the text document, or an entity not explicitly expressed in the document. The target toward which the sentiment is expressed, is either the frame or an entity in the document.

1) *Situation Frames*: Situation awareness information is encoded into situation frames in the LORELEI program [37]. Situation Frames (SF) are similar in nature to those used in Natural Language Understanding (NLU) systems: in essence they are data structures that record information corresponding to a single incident at a single location [16]. A SF frame includes a situation *Type* taken from a fixed inventory of 11 categories (e.g., medical need, shelter, infrastructure), *Location* where the situation exists (if a location is mentioned) and additional variables highlighting the *Status* of the situation (e.g., entities involved in resolution, time and urgency). An example of a SF can be found in table I. A list of situation frames and documents serve as input for our sentiment analysis systems.

### B. Data

Training data provided for the task included documents collected from social media, SMS, news articles, and news wires. This consisted of 76 documents in English and 47 in Spanish. All documents are relevant to the HADR domain but are not grounded in a common HADR incident. Each document is annotated for situation frames and associated sentiment by 2 trained annotators from the Linguistic Data

TABLE II  
 FREQUENCY STATISTICS FOR THE PROVIDED TRAINING DATA PER  
 LANGUAGE: NUMBER OF DOCUMENTS, NUMBER OF ANNOTATED  
 SITUATION FRAMES, NUMBER OF SENTIMENT INSTANCES, PERCENTAGE  
 OF NEGATIVE POLARITY.

	#Documents	#SF	#Sentiment	% Neg
English	76	85	380	81.57
Spanish	47	56	168	98.10
Total	123	141	548	84.85

Consortium (LDC)<sup>3</sup>. Sentiment annotations were done at a segment (sentence) level, and included Situation Frame, Polarity (positive / negative), Sentiment Score, Emotion, Source and Target. Sentiment labels were annotated between the values of -3 (very negative) and +3 (very positive) with 0.5 increments excluding 0. Additionally, the presence or absence of three specific emotions: fear, anger, and joy/happiness was marked. If a segment contains sentiment toward more than one target, each will be annotated separately. Summary of the training data is given in Table II.

### C. Evaluation

Systems participating in the task were expected to produce outputs with sentiment polarity, emotion, sentiment source and target, and the supporting segment from the input document. This output is evaluated against a ground truth derived from two or more annotations. For the SEC pilot evaluation, a reference set with dual annotations from two different annotators was provided. The system’s performance was measured using variants of precision, recall and f1 score, each modified to take into account the multiple annotations. The modified scoring is as follows: let the agreement between annotators be defined as two annotations with the same sentiment polarity, source, and target. That is, consider two annotators in agreement even if their judgments vary on sentiment values or perceived emotions. Designate those annotations with agreement as “D” and those which were not agreed upon as “S”. When computing precision, recall and f measure, each of the sentiment annotations in D will count as two occurrences in the reference, and likewise a system match on a sentiment annotation in D will count as two matches. Similarly, a match on a sentiment annotation in S will count as a single match. The updated precision, recall and f-measure were defined as follows:

$$\text{precision} = \frac{2 * \text{Matches in D} + \text{Matches in S}}{2 * \text{Matches in D} + \text{Matches in S} + \text{Unmatched}}$$

$$\text{recall} = \frac{2 * \text{Matches in D} + \text{Matches in S}}{2|D| + |S|}$$

$$\text{f1} = \frac{2 * \text{precision} * \text{recall}}{(\text{precision} + \text{recall})}$$

<sup>3</sup><https://www ldc upenn edu/>

## IV. METHOD

We approach the SEC task, particularly the polarity and emotion identification, as a classification problem. Our systems are based on English, and are extended to other languages via automatic machine translation (to English). In this section we present the linguistic features and describe the models using for the evaluation.

### A. Machine Translation

Automatic machine translation has been prove useful for previous years tasks of identifying Situation Frames and classifying their type [16], [36]. As a first step, we obtain translations from Spanish to English through the public API of Microsoft Bing<sup>4</sup>. For the pilot evaluation, we translated all of the Spanish documents into English, and included them as additional training data. At this time we do not translate English to Spanish, but plan to explore this thread in future work.

### B. Linguistic Features

1) *N-grams*: We extract word unigrams and bigrams. These features were then transformed using term frequencies (TF) and Inverse document-frequency (IDF).

2) *Distributed Semantics*: Word embeddings pretrained on large corpora allow models to efficiently leverage word semantics as well as similarities between words. This can help with vocabulary generalization as models can adapt to words not previously seen in training data. In our feature set we include a 300-dimensional word2vec word representation trained on a large news corpus [38]. We obtain a representation for each segment by averaging the embedding of each word in the segment. We also experimented with the use of GloVe [39], and Sent2Vec [40], an extension of word2vec for sentences.

3) *Sentiment Features*: We use two sources of sentiment features: manually constructed lexica, and pre-trained sentiment embeddings. When available, manually constructed lexica are a useful resource for identifying expressions of sentiment [22]. We obtained word percentages across 192 lexical categories using Empath [41], which extends popular tools such as the Linguistic Inquiry and Word Count (LIWC) [23] and General Inquirer (GI) [42] by adding a wider range of lexical categories. These categories include emotion classes such as surprise or disgust.

Neural networks have been shown to capture specific task related subtleties which can complement the manually constructed sentiment lexica described in the previous subsection. For this work, we learn sentiment representations using a bilateral Long Short-Term Memory model [43] trained on the Stanford Sentiment Treebank [44]. This model was selected because it provided a good trade off between simplicity and performance on a fine-grained sentiment task, and has been shown to achieve competitive results to the state-of-the-art [45].

<sup>4</sup><https://www bing com translator>

### C. Models

We now describe the models used for this work. Our models can be broken down into two approaches: our first approach explores *state-of-the-art* models in targeted and untargeted sentiment analysis to evaluate their performance in the context of the SEC task. These models were pre-trained on larger corpora and evaluated directly on the task without any further adaptation. In a second approach we explore a data augmentation technique based on a proposed simplification of the task. In this approach, traditional machine learning classifiers were trained to identify which segments contain sentiment towards a SF regardless of sentiment polarity. For the classifiers, we explored the use of Support Vector Machines and Random Forests. Model performance was estimated through 10-fold cross validation on the train set. Hyper-parameters, such as regularizations penalties, were selected based on the performance on grid-search using an 10-fold inner-cross validation loop. After choosing the parameters, models were re-trained on all the available data.

1) **Baselines:** We consider some of the most popular baseline models in the literature: (i) minority class baseline (due to the heavily imbalanced dataset), (ii) Support Vector Machines trained on TF-IDF bi-gram language model, (iii) and Support Vector Machines trained on word2vec representations. These models were trained using English documents only.

2) **Model I: Pretrained Sentiment Classifiers:** Two types of targeted sentiment are evaluated for the task: those expressed towards either a situation frame or those towards an entity. To identify sentiment expressed towards an SF, we use the pretrained model described in [46], in which a multiplicative LSTM cell is trained at the character level on a corpus of 82 million Amazon reviews. The model representation is then fed to a logistic regression classifier to predict sentiment. This model (which we will refer to as OpenAI) was chosen since at the time of our system submission it was one of the top three performers on the binary sentiment classification task on the Stanford Sentiment Treebank. In our approach, we first map the text associated with the SF annotation with a segment from the document and pass the full segment to the pretrained OpenAI model identify the sentiment polarity for that segment.

To identify sentiment targeted towards an entity, we use the recently released Target-Based Sentiment Analysis (TBSA) model from [47]. In TBSA, two stacked LSTM cells are trained to predict both sentiment and target boundary tags (e.g., predicting S-POS to indicate the start of the target towards which the author is expressing positive sentiment, I-POS and E-POS to indicate intermediate and end of the target). In our submission, since input text documents can be arbitrarily long, we only consider sentences which include a known and relevant entity; these segments are then fed to the TBSA model to predict targeted sentiment. If the target predicted by this model matched with any of the known entities, the system would output the polarity and the target.

3) **Model IIa: Simplifying the Task:** In this model we limit our focus on the task of correctly identifying those segments with sentiment towards a SF. That is, given a pair of SF and

TABLE III  
FREQUENCY STATISTICS FOR THE TRAIN DATASET AFTER AUGMENTATION

	#(SF×Segments)	With Sentiment	Total
English	5751	285	6030
Spanish	1232	132	1364

TABLE IV  
TRAIN DATASET DOMAIN BREAK-DOWN

English	Neg	Pos	Spanish	Neg	Pos
Tweet	85	16	Tweet	47	1
Others	204	43	Others	98	12
Total	289	59	Total	145	13

segment, we train models to identify if this segment contains any sentiment towards that SF. This allows us to expand our dataset from 123 documents into one with  $\sum_d |SF_d| \times |d|$  number of samples, where  $|d|$  is the length of the document (i.e., number of segments) and  $|SF_d|$  is the number of SF annotations for document  $d$ . Summary of the training dataset after augmentation is given in Table III.

Given the highly skewed label distribution in the training data, a majority of the constructed pairs do not have any sentiment towards a SF. Hence, our resulting dataset has a highly imbalanced distribution. We control for this imbalance by setting class' weights to be the inverse of the class frequency during train. To predict polarity, we assume the majority class (i.e., negative sentiment). We base this assumption after initial experiments revealed that there were not many instances for positive sentiment, which can be attributed both to the domain (HADR) and the source of the messages (e.g., Twitter).

4) **Model IIb: Domain-specific models:** Following our discoveries in domain and source differences, in a second approach we train two different SVM models: one for tweets and one for non-tweets. This approach aims to capture the underlying difference in the way sentiment is expressed across social media and other types of media [27]. Moreover, this also addresses the considerable variance in the structure of the documents, since tweets only have one segment per sample whereas news articles contain an average of  $7.07 \pm 4.96$  and  $6.31 \pm 4.93$  segments for English and Spanish documents respectively. Table IV presents a breakdown of the train set for each sentiment across domains.

5) **Model IIc: Twitter-only model:** Initial experiments showed that our main source of error was not being able to correctly identify the supporting segment. Even if polarity, source and target were correctly identified, missing the correct segment was considered an error, and thus lowered our models' precision. To address this, we decided to only use the output of the tweet-model from models IIb given that tweets only contain one segment (making the segment identification sub-task trivial). While this model will perform poorly in recall, we will see that the gains in precision more than make up for the difference.

TABLE V  
MODEL PERFORMANCE ON ENGLISH TRAIN DATA ESTIMATED USING  
10-FOLD CV

Model	Prec	Recall	F1
Minority	0.04	<b>1.00</b>	0.08
SVM tfidf	0.69	0.12	0.21
SVM W2V	0.10	0.38	0.16
Model IIa	0.42	0.17	0.24
Model IIb	0.92	0.22	<b>0.36</b>
Model IIc	<b>1.00</b>	0.22	<b>0.36</b>

TABLE VI  
OFFICIAL EVALUATION RESULTS FOR ENGLISH AND SPANISH. DASHES  
DENOTE MISSING INFORMATION (NOT REPORTED)

Polarity	Eng	Spa
Team 1	0.33	0.02
Team 2	0.03	0.04
Model I	0.20	-
Model IIa	0.03	0.05
Model IIb	0.32	0.35
Model IIc	<b>0.36</b>	<b>0.39</b>

## V. RESULTS

Model performance during train is presented in Table V. While all the models outperformed the baselines, not all of them did so with a significant margin due to the robustness of the baselines selected. The ones found to be significantly better than the baselines were models IIb (Domain-specific) and IIc (Twitter-only) (permutation test,  $n = 10^5$  both  $p < 0.05$ ). The difference in precision between model IIb and IIc points out to the former making the wrong predictions for news articles. These errors are most likely in selecting the wrong supporting segment. Moreover, even though models IIa-c only produce negative labels, they still achieve improved performance over the state-of-the-art systems, highlighting the highly skewed nature of the training dataset.

Table VI present the official evaluation results for English and Spanish. Some information is missing since at the time of submission only partial score had been made public. As previously mentioned, the pre-trained *state-of-the-art* models (model I) were directly applied to the evaluation data without any adaptation. These performed reasonably well for the English data. Among the submissions of the SEC Task pilot, our systems outperformed the other competitors for both languages.

## VI. CONCLUSION

Understanding the expressed sentiment from an affected population during the on-set of a crisis is a particularly difficult task, especially in low-resource scenarios. There are multiple difficulties beyond the limited amount of data. For example, in order to provide decision-makers with actionable and usable information, it is not enough for the system to correctly classify sentiment or emotional state, it also ought to identify the source and target of the expressed sentiment. To provide a sense of trust and accountability on the system’s decisions, it makes sense to identify a justifying segment. Moreover, these

systems should consider a variety of information sources to create a broader and richer picture on how a situation unfolds. Thus, it is important that systems take into account the possible differences in the way sentiment is expressed in each one of these sources. In this work, we presented two approaches to the task of providing actionable and useful information. Our results show that *state-of-the-art* sentiment classifiers can be leveraged out-of-the-box for a reasonable performance on English data. By identifying possible differences coming from the information sources, as well as by exploiting the information communicated as the situation unfolds, we showed significant performance gains on both English and Spanish.

## VII. ACKNOWLEDGEMENTS

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0115 with the University of Southern California. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA.

NIST serves to coordinate the evaluations in order to support research and to help advance the state-of-the-art. NIST evaluations are not viewed as a competition, and such results reported by NIST are not to be construed, or represented, as endorsements of any participants system, or as official findings on the part of NIST or the U.S. Government.

## REFERENCES

- [1] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, “Processing social media messages in mass emergency: A survey,” *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, p. 67, 2015.
- [2] A. L. Hughes and L. Palen, “Twitter adoption and use in mass convergence and emergency events,” *International journal of emergency management*, vol. 6, no. 3-4, pp. 248–260, 2009.
- [3] S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen, “Microblogging during two natural hazards events: what twitter may contribute to situational awareness,” in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 1079–1088.
- [4] S. Vieweg, “Twitter communications in mass emergency: contributions to situational awareness,” in *Proceedings of the ACM 2012 conference on computer supported cooperative work companion*. ACM, 2012, pp. 227–230.
- [5] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, “Aidr: Artificial intelligence for disaster response,” in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 159–162.
- [6] S. Vieweg, C. Castillo, and M. Imran, “Integrating social media communications into the rapid assessment of sudden onset disasters,” in *International Conference on Social Informatics*. Springer, 2014, pp. 444–461.
- [7] J. R. Harrald, D. M. Egan, T. Jefferson, E. Stok, and B. Žmavc, “Web enabled disaster and crisis response: What have we learned from the september 11 th,” *Proceedings of the Bled eConference*, pp. 69–83, 2002.
- [8] M. Imran, P. Mitra, and C. Castillo, “Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, 2016.
- [9] R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt, “Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events,” *Computers in Human Behavior*, vol. 56, pp. 179–191, 2016.
- [10] H. Purohit, A. Hampton, V. L. Shalin, A. P. Sheth, J. Flach, and S. Bhatt, “What kind of# conversation is twitter? mining# psycholinguistic cues for emergency coordination,” *Computers in Human Behavior*, vol. 29, no. 6, pp. 2438–2447, 2013.

- [11] K. A. Lachlan, P. R. Spence, and X. Lin, "Expressions of risk awareness and concern through twitter: On the utility of using the medium as an indication of audience needs," *Computers in Human Behavior*, vol. 35, pp. 554–559, 2014.
- [12] M. K. Torkildson, K. Starbird, and C. Aragon, "Analysis and visualization of sentiment and emotion on crisis tweets," in *International Conference on Cooperative Design, Visualization and Engineering*. Springer, 2014, pp. 64–67.
- [13] A. L. Hughes and L. Palen, "The evolving role of the public information officer: An examination of social media in emergency management," *Journal of Homeland Security and Emergency Management*, vol. 9, no. 1, 2012.
- [14] A. L. Hughes, L. A. St Denis, L. Palen, and K. M. Anderson, "Online public communications by police & fire services during the 2012 hurricane sandy," in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 1505–1514.
- [15] L. L. Cheung, T. Gowda, U. Hermjakob, N. H. S. Liu, J. May, A. Mayn, N. Pourdamghani, M. Pust, K. Knight, N. Malandrakis, P. Papadopoulos, A. Ramakrishna, K. Singla, V. C. Martínez, C. Vaz, D. Can, S. S. Narayanan, K. Murray, T. Nguyễn, D. Chiang, X. Pan, B. Zhang, Y. C. Lin, D. Lu, L. Huang, K. Blissett, T. Zhang, O. Glembek, M. K. Baskar, S. Kesiraju, L. Burget, K. Benes, I. Szoke, K. Veselý, C. Goudeseune, M. H. Johnson, L. Sari, W. Chen, and A. Liu, "Elisa system description for lorehlt 2017," 2017.
- [16] N. Malandrakis, A. Ramakrishna, V. Martinez, T. Sorensen, D. Can, and S. Narayanan, "The elisa situation frame extraction for low resource languages pipeline for lorehlt'2016," *Machine Translation*, vol. 32, no. 1, pp. 127–142, Jun 2018.
- [17] A. Schulz, T. D. Thanh, H. Paulheim, and I. Schweizer, "A fine-grained sentiment analysis approach for detecting crisis related microposts," in *ISCRAM*, 2013.
- [18] I. Gilles, A. Bangerter, A. Clémence, E. G. Green, F. Krings, A. Mouton, D. Rigaud, C. Staerklé, and P. Wagner-Egger, "Collective symbolic coping with disease threat and othering: A case study of avian influenza," *British Journal of Social Psychology*, vol. 52, no. 1, pp. 83–102, 2013.
- [19] D. Murthy and S. A. Longwell, "Twitter and disasters: The uses of twitter during the 2010 pakistan floods," *Information, Communication & Society*, vol. 16, no. 6, pp. 837–855, 2013.
- [20] P. Panagiotopoulos, A. Z. Bigdeli, and S. Sams, "Citizen–government collaboration on social media: The case of twitter in the 2011 riots in england," *Government information quarterly*, vol. 31, no. 3, pp. 349–357, 2014.
- [21] G. Neubaum, L. Rösner, A. M. Rosenthal-von der Pütten, and N. C. Krämer, "Psychosocial functions of social media usage in a disaster situation: A multi-methodological approach," *Computers in Human Behavior*, vol. 34, pp. 28–38, 2014.
- [22] G. Beigi, X. Hu, R. Maciejewski, and H. Liu, "An overview of sentiment analysis in social media and its applications in disaster relief," in *Sentiment analysis and ontology engineering*. Springer, 2016, pp. 313–340.
- [23] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.
- [24] F. Å. Nielsen, "Afinn," Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby, mar 2011.
- [25] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," in *LREC*, vol. 6. Citeseer, 2006, pp. 417–422.
- [26] J. Brynielsson, F. Johansson, and A. Westling, "Learning to classify emotional content in crisis-related tweets," in *2013 IEEE International Conference on Intelligence and Security Informatics*. IEEE, 2013, pp. 33–38.
- [27] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-scale sentiment analysis for news and blogs," *Icwsn*, vol. 7, no. 21, pp. 219–222, 2007.
- [28] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, and J. Belyaeva, "Sentiment analysis in the news," *arXiv preprint arXiv:1309.6202*, 2013.
- [29] M. Kaya, G. Fidan, and I. H. Toroslu, "Sentiment analysis of turkish political news," in *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 01*. IEEE Computer Society, 2012, pp. 174–180.
- [30] E. J. De Fortuny, T. De Smedt, D. Martens, and W. Daelemans, "Media coverage in times of political crisis: A text mining approach," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11 616–11 622, 2012.
- [31] N. Öztürk and S. Ayvaz, "Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis," *Telematics and Informatics*, vol. 35, no. 1, pp. 136–147, 2018.
- [32] A. Zielinski, U. Bügel, L. Middleton, S. Middleton, L. Tokarchuk, K. Watson, and F. Chaves, "Multilingual analysis of twitter news in support of mass emergency events," in *EGU General Assembly Conference Abstracts*, vol. 14. Citeseer, 2012, p. 8085.
- [33] G. Neubig, Y. Matsubayashi, M. Hagiwara, and K. Murakami, "Safety information mining: what can nlp do in a disaster," in *Proceedings of 5th International Joint Conference on Natural Language Processing*, 2011, pp. 965–973.
- [34] G. Shalunts and G. Backfried, "Sentsail: sentiment analysis in english, german and russian," in *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, 2015, pp. 87–97.
- [35] A. Zielinski, "Detecting natural disaster events on twitter across languages," in *IHMSS*, 2013, pp. 291–301.
- [36] P. Papadopoulos, R. Travadi, C. Vaz, N. Malandrakis, U. Hermjakob, N. Pourdamghani, M. Pust, B. Zhang, X. Pan, D. Lu *et al.*, "Team elisa system for darpa lorelei speech evaluation 2016," in *INTERSPEECH*, 2017, pp. 2053–2057.
- [37] S. M. Strassel and J. Tracey, "Lorelei language packs: Data, tools, and resources for technology development in low resource languages," in *LREC*, 2016.
- [38] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.
- [39] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.
- [40] M. Pagliardini, P. Gupta, and M. Jaggi, "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*, 2018.
- [41] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 4647–4657.
- [42] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis," 1966.
- [43] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [44] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.
- [45] J. Barnes, R. Klinger, and S. Schulte im Walde, "Assessing state-of-the-art sentiment models on state-of-the-art sentiment datasets," in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, 2017, pp. 2–12.
- [46] A. Radford, R. Józefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *CoRR*, vol. abs/1704.01444, 2017. [Online]. Available: <http://arxiv.org/abs/1704.01444>
- [47] X. Li, L. Bing, P. Li, and W. Lam, "A unified model for opinion target extraction and target sentiment prediction," *arXiv preprint arXiv:1811.05082*, 2018.