IJCNLP-2017 Task 4: Customer Feedback Analysis

Chao-Hong Liu, Yasufumi Moriya, Alberto Poncelas
ADAPT Centre, Ireland
Dublin City University
{chaohong.liu, yasufumi.moriya, alberto.poncelas}@adaptcentre.ie

Declan Groves
Microsoft Ireland
Leopardstown, Dublin 18
degroves @microsoft.com

Abstract

This document introduces the IJCNLP 2017 Shared Task on Customer Feedback Analysis. In this shared task we have prepared corpora of customer feedback in four languages, i.e. English, French, Spanish and Japanese. They were annotated in a common meanings categorization, which was improved from an ADAPT-Microsoft pivot study on customer feedback. Twenty teams participated in the shared task and twelve of them have submitted prediction results. The results show that performance of prediction meanings of customer feedback is reasonable well in four languages. Nine system description papers are archived in the shared tasks proceeding.

1 Introduction

In this paper we introduce the results of IJCNLP 2017 Shared Task on Customer Feedback Analysis. The shared task is a follow-up of an ADAPT-Microsoft joint pilot study on multilingual customer feedback analysis. We have improved the categorization and the classes (tags) used in the corpora are the five-class “comment”, “request”, “bug”, “complaint”, “meaningless”, and the “undetermined” tag. By undetermined we mean that the feedback could be annotated as one of the five classes but due to lack of contexts it was annotated as undetermined. Table 1 shows the numbers of customer feedback sentences curated in the corpora and how many they are grouped into training, development and test sets. We also provided unannotated customer feedback sentences in the corpora. Table 2 shows the statistics of each class in the meaning categorization in the training set. Noted we cannot find “meaningless” feedback sentence in Japanese corpus. On the contrary, there is no “undetermined” feedback sentence in Spanish corpus. These might reflect some linguistic and/or cultural differences in the curated customer feedback corpora. Abbreviations EN, ES, FR and JP are used interchangeably with English, Spanish, French and Japanese where applicable.

Lang.	Train.	Dev.	Test	Unanno.
English	3,065	500	500	12,838
French	1,950	400	400	5,092
Spanish	1,631	301	299	6,035
Japanese	1,526	250	300	4,873
TOTAL	8,172	1,451	1,499	28,838

Table 1: Statistics of the curated Customer Feedback Analysis Corpora for the shared task.

	EN	FR	ES	JP
Comment	276	259	224	142
Request	21	6	12	22
Bug	21	13	5	18
Complaint	148	112	39	73
Meaningless	48	36	1	0
Undetermined	3	1	0	9

Table 2: Numbers of customer feedback tags that were annotated in the training set.

The purpose of the shared task is to try to answer the question that if we need to 1) train native systems for different languages (using the same meanings categorization of customer feedback), or it is good enough to 2) use Machine Translation (MT) to translate customer feedback in other languages into English and use English based systems to do the detection of meanings of customer feedback. If the answer is 1, we will have to prepare corpora for different languages using the same categorization. If the answer is 2, then it would be more reasonable to put more efforts to

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 26–33,
Taipei, Taiwan, November 27 – December 1, 2017.

enhance the performance of English based systems and try to further improve the quality of MT results.

There are several categorizations that could be used for customer feedback analysis. First, different kinds of sentiment categorizations that were used in sentiment analysis in Microsoft Office and many other institutions (Salameh et al., 2015) Customer feedback analysis is now an industry in its own right (Freshdesk, 2016; Burns, 2016). One commonly used categorization is the Excellent-Good-Average-Fair-Poor and its various kinds of variants (Yin et al., 2016; SurveyMonkey, 2016). (Freshdesk, 2016) and (Keatext, 2016) used a combined categorization of Positive-Neutral-Negative-Answered-Unanswered. (Sift, 2016) has the Refund-Complaint-Pricing-Tech Support-Store Locator-Feedback-Warranty Info categorization in seven classes. We can also have observed that there are many other categorizations that are not publicly available (Equiniti, 2016; UseResponse, 2016; Inmoment, 2016).

In this shared task, we followed (Liu et al., 2017)'s five-class customer feedback meanings categorization which is generalized from English, Spanish and Japanese customer feedback, add an "undetermined" class and prepared the corpora in four languages (English, French, Spanish and Japanese). The resulting categorization is as follows.

1. Comment
2. Request
3. Bug
4. Complaint
5. Meaningless
6. Undetermined

2 Measures

In this shared task, we concluded the results in four different measures. The details of the results can be download from the shared task website.

- Exact-match Accuracy: Feedback is considered correct only when "all its oracle tags" are predicted correctly.
- Partial-match Accuracy: Feedback is considered correct if 'any' of its oracle tags is predicted.
- Micro-Average of Precision, Recall and F1
- Macro-Average of Precision, Recall and F1: As the number of instances of each tag varies a lot this measure might not be suitable for comparisons in the shared task.

In this paper we show mainly the results of 1) Exact-match Accuracy and 2) Micro-Average of Precision, Recall and F1, which are more suitable measures in our consideration.

3 Baseline and Submitted Systems

A baseline system was implemented using similarity based method. It uses trigrams to calculate the similarity of an input sentence and all the annotated customer feedback sentences in the corpora and uses the annotation of the one (in the annotated training corpora) with highest similarity score as the input sentence's predicted annotation. The baseline system is referred to as "Baseline-Similarity" in this paper.

In this shared task, an initial team name was given to each team in the release of results. For example, TA was used to designate Team A. In the report of these results, i.e. this paper, a team name is revealed only when consent from its corresponding team is granted.

The mapping of each team name and its corresponding system description paper is shown as follows. Please refer to each paper for details of the system/method they used for the problem of customer feedback analysis

- ADAPT: (Lohar et al., 2017)
- Bingo: (Elfardy et al., 2017)
- IIIT-H: (Danda et al., 2017)
- OhioState: (Dhyani, 2017)
- Plank: (Plank, 2017)
- SentiNLP: (Lin et al., 2017)
- YNU-HPCC: (Wang et al., 2017)

4 Results in Exact-Match Accuracy

Tables 3-6 shows the results of each team-method in exact-match accuracy in English, Spanish, French and Japanese, respectively. The details of each method implemented by each team are described in their associated system description papers. The method denoted as "entrans" is the one that used machine translated sentences to do the prediction of meanings of customer feedback. For example, in the "Plank-entrans" system in Table 4, the sentences in Spanish test set are machine translated from Spanish to English using Google Translate, and then use Plank's English based system to predict their tags.

It is observed that for exact-accuracy, the best performers of submitted systems can achieve 71.00%, 88.63%, 73.75% and 75.00% in English, Spanish, French and Japanese, respectively. First, we can observe that the task seems to be easier in

Spanish which is the same phenomenon reported in (Liu et al., 2017). Second, performances in English, French and Japanese are also good and around the same level. Third, using machine translation the systems can achieve comparable results for Spanish and French, which are only 4 and 2 points behind native systems, respectively. For Japanese there is about 12 points behind the best native system.

English	Exact-Accuracy
YNU-HPCC-glove	**71.00%**
YNU-HPCC-EmbedConcatNoWeight	71.00%
SentiNLP-bilstmcnn	70.80%
SentiNLP-bilstm	70.40%
SentiNLP-bicnn	70.20%
IITP-CNN	70.00%
SentiNLP-cnnlstm	69.00%
Plank-monolingual	68.80%
Plank-multilingual	68.60%
YNU-HPCC-EmbedConcatWeight	68.60%
SentiNLP-cnn	68.20%
TJ-single-cnn	67.40%
IIIT-H-SVM	65.60%
TJ-ensemble-sentiment	65.40%
ADAPT-Run3	65.40%
IIIT-H-biLSTM	65.20%
TJ-ensemble-2	65.20%
YNU-HPCC-hotelWeight	65.00%
TJ-ensemble-epoch5	64.60%
TJ-ensemble-7	64.60%
TJ-ensemble-1	64.60%
TJ-ensemble-epoch10	64.40%
TJ-ensemble-5	64.20%
YNU-HPCC-hotel	64.00%
YNU-HPCC-gloveWeight	64.00%
TJ-ensemble-epoch5n10	64.00%
ADAPT-Run2	64.00%
TJ-ensemble-8	63.80%
TJ-ensemble-6	63.80%
TJ-ensemble-3	63.80%
TJ-ensemble-4	63.60%
OhioState-FastText	63.40%
ADAPT-Run1	63.40%
YNU-HPCC-SVM	63.00%
OhioState-biLSTM3	62.80%
YNU-HPCC-bayes	62.60%
TJ-single-cbow	62.00%
OhioState-biLSTM2	61.60%
IITP-RNN	61.40%
YNU-HPCC-hotelNoATT	61.20%
OhioState-biLSTM1	61.20%
Bingo-logistic-reg	55.80%
Bingo-lstm	54.40%
OhioState-CNN	54.20%
TD-M1	52.20%
TF-nn	51.20%
Baseline-Similarity	**48.80%**
Bingo-rf	47.40%
TF-ss-svm	41.00%
TF-ss-lr	41.00%
TF-ss-nb	40.40%
TF-ss	40.40%
TB-en-run2	38.80%
TB-en-run1	37.40%
TB-en-run3	37.00%

Table 3: Resulting scores of each team-method in exact-match accuracy in English.

Spanish	Exact-Accuracy
Plank-multilingual	**88.63%**
Plank-monolingual	88.29%
IIIT-H-biLSTM	86.29%
IITP-RNN	85.62%
OhioState-biLSTM2	85.28%
Plank-entrans	84.62%
IITP-CNN	84.62%
IIIT-H-SVM	84.62%
ADAPT-Run1	83.61%
OhioState-FastText	82.94%
IITP-CNN-entrans	82.61%
OhioState-biLSTM1	82.61%
IITP-RNN-entrans	81.94%
ADAPT-Run2	81.61%
OhioState-CNN	81.27%
OhioState-biLSTM3	79.93%
Baseline-Similarity	**77.26%**
TF-ss-lr-entrans	76.25%
Bingo-rf	75.92%
Bingo-logistic-reg	72.91%
Bingo-lstm	71.57%
TF-ss	62.21%
TF-cnn-entrans	60.54%
TF-nn	59.53%
TF-ss-svm	57.19%
TF-ss-nb	57.19%

TF-ss-lr	57.19%

Table 4: Resulting scores of each team-method in exact-match accuracy in Spanish.

French	Exact-Accuracy
Plank-monolingual	**73.75%**
IITP-CNN-entrans	71.75%
Plank-multilingual	71.50%
OhioState-biLSTM1	70.00%
IIIT-H-SVM	69.75%
ADAPT-Run1	69.50%
IITP-CNN	69.00%
OhioState-biLSTM2	68.50%
Plank-entrans	68.25%
IITP-RNN-entrans	68.25%
IITP-RNN	68.25%
OhioState-FastText	68.00%
TB-fr-run1	66.75%
ADAPT-Run2	66.75%
IIIT-H-biLSTM	65.25%
OhioState-biLSTM3	65.00%
OhioState-CNN	65.00%
TB-fr-run4	63.50%
TB-fr-run3	62.25%
Bingo-lstm	61.25%
TB-fr-run2	60.50%
Bingo-logistic-reg	59.00%
Baseline-Similarity	**54.75%**
Bingo-rf	48.75%
TF-ss-nb	48.25%
TF-ss-lr	48.25%
TF-nn2	47.75%
TF-nn	47.25%
TF-ss	44.50%
TF-nn3	39.00%

Table 5: Resulting scores of each team-method in exact-match accuracy in French.

Japanese	Exact-Accuracy
Plank-multilingual	**75.00%**
Plank-monolingual	73.33%
ADAPT-Run1	67.67%
Plank-entrans	63.67%
IITP-CNN-entrans	63.00%
Bingo-logistic-reg	60.67%
IITP-RNN-entrans	58.67%
ADAPT-Run2	57.67%
Baseline-Similarity	**56.67%**
IIIT-H-biLSTM	56.67%
OhioState-CNN	56.67%
IIIT-H-SVM	56.33%
OhioState-biLSTM1	56.33%
OhioState-biLSTM2	56.33%
IITP-RNN	56.00%
OhioState-biLSTM3	56.00%
OhioState-FastText	56.00%
TF-nn	55.67%
TF-ss-svm	55.00%
TF-ss-nb	55.00%
TF-ss	55.00%
IITP-CNN	54.00%
TF-cnn-entrans	53.33%
TF-ss-lr-entrans	53.00%
Bingo-lstm	53.00%
Bingo-rf	45.00%
TF-ss-lr	28.67%

Table 6: Resulting scores of each team-method in exact-match accuracy in Japanese.

5 Results in Micro-Average Precision, Recall and F1 measures

Likewise, Tables 7-10 show the results of each team-method in micro-average precision, recall and F1 measures in English, Spanish, French and Japanese, respectively.

For micro-average F1, the best systems achieved 75.57%, 88.63%, 76.59% and 77.05% in English, Spanish, French and Japanese, respectively. The results in Spanish exhibit the same phenomenon as in exact-match accuracy results and in (Liu et al., 2017). The performances in English, French and Japanese are also good and around the same level. Using machine translation, the systems can also achieve comparable results in this measure for Spanish and French, which are 4 and 2 points behind native systems, respectively. There is 11 points behind in Japanese in this regard.

English	A.P	A.R	A.F1
SentiNLP-bilstmcnn	**74.86%**	**76.30%**	**75.57%**
SentiNLP-bicnn	73.83%	76.11%	74.95%
SentiNLP-bilstm	73.77%	75.34%	74.55%
SentiNLP-cnn	72.12%	74.76%	73.42%

SentiNLP-cnnlstm	73.04%	73.60%	73.32%
YNU-HPCC-EmbedConcat-NoWeight	74.60%	71.87%	73.21%
YNU-HPCC-glove	74.40%	71.68%	73.01%
IITP-CNN	73.80%	71.10%	72.42%
Plank-monolingual	72.40%	69.75%	71.05%
Plank-multilingual	72.20%	69.56%	70.85%
TJ-single-cnn	70.85%	70.71%	70.78%
YNU-HPCC-EmbedConcat-Weight	72.00%	69.36%	70.66%
TJ-ensemble-2	69.10%	69.36%	69.23%
TJ-ensemble-sentiment	68.22%	67.82%	68.02%
TJ-ensemble-epoch10	68.02%	68.02%	68.02%
TJ-ensemble-8	67.82%	68.21%	68.01%
TJ-ensemble-epoch5n10	66.73%	69.17%	67.93%
ADAPT-Run3	69.20%	66.67%	67.91%
IIIT-H-biLSTM	67.82%	67.82%	67.82%
TJ-ensemble-epoch5	67.62%	68.02%	67.82%
IIIT-H-SVM	69.22%	66.28%	67.72%
YNU-HPCC-hotelWeight	68.60%	66.09%	67.32%
TJ-ensemble-1	67.77%	66.86%	67.31%
TJ-ensemble-4	67.18%	67.44%	67.31%
TJ-ensemble-6	66.92%	67.44%	67.18%
TJ-ensemble-7	67.18%	67.05%	67.12%
TJ-ensemble-3	66.79%	67.44%	67.11%
TJ-ensemble-5	66.99%	66.47%	66.73%
YNU-HPCC-hotel	67.60%	65.13%	66.34%
YNU-HPCC-gloveWeight	67.40%	64.93%	66.14%
ADAPT-Run2	67.40%	64.93%	66.14%
ADAPT-Run1	66.87%	64.55%	65.69%
YNU-HPCC-SVM	66.60%	64.16%	65.36%
OhioState-FastText	66.60%	64.16%	65.36%
OhioState-biLSTM3	66.20%	63.78%	64.97%
YNU-HPCC-bayes	66.00%	63.58%	64.77%
TJ-single-cbow	65.11%	64.35%	64.73%
Bingo-rf	54.35%	79.38%	64.53%
OhioState-biLSTM2	65.20%	62.81%	63.98%
OhioState-biLSTM1	65.00%	62.62%	63.79%
IITP-RNN	64.60%	62.24%	63.40%
YNU-HPCC-hotelNoATT	63.40%	61.08%	62.22%
Bingo-logistic-reg	60.47%	63.97%	62.17%
TD-M1	55.78%	68.79%	61.60%
Bingo-lstm	56.97%	65.32%	60.86%
OhioState-CNN	57.20%	55.11%	56.13%
Baseline-Similarity	**53.73%**	**54.14%**	**53.93%**
TF-nn	54.40%	52.41%	53.39%
TB-en-run1	42.70%	44.51%	43.58%
TF-ss-svm	44.20%	42.58%	43.38%
TF-ss-lr	44.20%	42.58%	43.38%
TF-ss-nb	43.60%	42.00%	42.79%
TF-ss	43.60%	42.00%	42.79%
TB-en-run2	42.77%	42.77%	42.77%
TB-en-run3	41.32%	42.20%	41.75%

Table 7: Resulting scores of each team-method in micro-average precision (A.P), recall (A.R) and F1 (A.F1) measures in English.

Spanish	A.P	A.R	A.F1
Plank-multilingual	**88.63%**	**88.63%**	**88.63%**
Plank-monolingual	88.29%	88.29%	88.29%

IIIT-H-biLSTM	86.29%	86.29%	86.29%
IITP-RNN	85.62%	85.62%	85.62%
OhioState-biLSTM2	85.28%	85.28%	85.28%
Plank-en-trans	84.62%	84.62%	84.62%
IITP-CNN	84.62%	84.62%	84.62%
IIIT-H-SVM	84.62%	84.62%	84.62%
ADAPT-Run1	83.61%	83.61%	83.61%
OhioState-FastText	82.94%	82.94%	82.94%
IITP-CNN-entrans	82.61%	82.61%	82.61%
OhioState-biLSTM1	82.61%	82.61%	82.61%
IITP-RNN-entrans	81.94%	81.94%	81.94%
ADAPT-Run2	81.61%	81.61%	81.61%
OhioState-CNN	81.27%	81.27%	81.27%
Bingo-logistic-reg	82.29%	79.26%	80.75%
OhioState-biLSTM3	79.93%	79.93%	79.93%
Bingo-lstm	71.35%	86.62%	78.25%
Bingo-rf	75.00%	81.27%	78.01%
Baseline-Similarity	**77.26%**	**77.26%**	**77.26%**
TF-ss-lr-en-trans	76.25%	76.25%	76.25%
TF-ss	62.21%	62.21%	62.21%
TF-cnn-en-trans	60.54%	60.54%	60.54%
TF-nn	59.53%	59.53%	59.53%
TF-ss-svm	57.19%	57.19%	57.19%
TF-ss-nb	57.19%	57.19%	57.19%
TF-ss-lr	57.19%	57.19%	57.19%

Table 8: Resulting scores of each team-method in micro-average precision (A.P), recall (A.R) and F1 (A.F1) measures in Spanish.

French	A.P	A.R	A.F1
Plank-monolingual	**78.50%**	**74.76%**	**76.59%**
IITP-CNN-entrans	76.50%	72.86%	74.63%
Plank-multilingual	76.25%	72.62%	74.39%
OhioState-biLSTM1	75.00%	71.43%	73.17%
ADAPT-Run1	74.50%	70.95%	72.68%
IIIT-H-SVM	74.68%	70.24%	72.39%
IITP-CNN	73.50%	70.00%	71.71%
OhioState-biLSTM2	73.50%	70.00%	71.71%
Plank-en-trans	73.00%	69.52%	71.22%
IITP-RNN	72.75%	69.29%	70.98%
IITP-RNN-entrans	72.25%	68.81%	70.49%
OhioState-FastText	72.25%	68.81%	70.49%
TB-fr-run1	70.94%	69.76%	70.35%
Bingo-lstm	62.04%	79.76%	69.79%
ADAPT-Run2	71.00%	67.62%	69.27%
IIIT-H-biLSTM	72.63%	65.71%	69.00%
TB-fr-run4	68.10%	68.10%	68.10%
OhioState-CNN	69.50%	66.19%	67.80%
OhioState-biLSTM3	69.25%	65.95%	67.56%
Bingo-rf	58.03%	80.00%	67.27%
TB-fr-run3	66.75%	65.95%	66.35%
Bingo-logistic-reg	64.84%	67.62%	66.20%
TB-fr-run2	65.02%	62.86%	63.92%
Baseline-Similarity	**60.05%**	**60.48%**	**60.26%**
TF-nn3	56.48%	53.80%	55.11%
TF-ss-nb	52.25%	49.76%	50.98%
TF-ss-lr	52.25%	49.76%	50.98%
TF-nn2	51.75%	49.29%	50.49%
TF-nn	51.50%	49.05%	50.24%
TF-ss	48.75%	46.43%	47.56%

Table 9: Resulting scores of each team-method in micro-average precision (A.P), recall (A.R) and F1 (A.F1) measures in French.

Japanese	A.P	A.R	A.F1
Plank-multilingual	**79.12%**	**75.08%**	**77.05%**
Plank-monolingual	77.70%	73.48%	75.53%
ADAPT-Run1	71.67%	68.69%	70.15%
IITP-CNN-entrans	70.21%	63.26%	66.55%
Bingo-rf	56.36%	79.23%	65.87%

Plank-en-trans	67.46%	63.58%	65.46%
Bingo-logistic-reg	63.86%	65.50%	64.67%
Bingo-lstm	58.38%	71.25%	64.17%
IITP-RNN-entrans	65.60%	59.11%	62.18%
ADAPT-Run2	62.24%	58.47%	60.30%
Baseline-Similarity	**59.24%**	**59.42%**	**59.33%**
OhioState-CNN	58.00%	55.59%	56.77%
IIIT-H-biLSTM	57.33%	54.95%	56.12%
IIIT-H-SVM	57.33%	54.95%	56.12%
OhioState-biLSTM1	57.00%	54.63%	55.79%
OhioState-biLSTM2	57.00%	54.63%	55.79%
IITP-RNN	56.67%	54.31%	55.46%
OhioState-biLSTM3	56.67%	54.31%	55.46%
OhioState-FastText	56.67%	54.31%	55.46%
TF-nn	56.33%	53.99%	55.14%
TF-ss-svm	55.67%	53.35%	54.49%
TF-ss-nb	55.67%	53.35%	54.49%
TF-ss	55.67%	53.35%	54.49%
IITP-CNN	55.00%	52.72%	53.83%
TF-cnn-en-trans	54.67%	52.40%	53.51%
TF-ss-lr-en-trans	53.67%	51.44%	52.53%
TF-ss-lr	32.67%	31.31%	31.97%

Table 10: Resulting scores of each team-method in micro-average precision (A.P), recall (A.R) and F1 (A.F1) measures in Japanese.

6 Conclusions

In this shared task, we address the problem if we should 1) train native systems for different languages, or 2) use MT to translate customer feedback into English and use English based systems to predict meanings of customer feedback. By using the same categorization, we concluded that using native systems, the performances in the four languages are all good. For Spanish and French, using MT can achieve comparable results as using native systems. Therefore, we would suggest improving English based systems and probably preparing the corpora in finer categorizations that would help us understand customer feedbacks. However, for Japanese or other languages where MT still does not produce high quality translations, preparing native corpora and building native systems are still highly recommended.

Acknowledgements

This research is supported by the ADAPT Centre for Digital Content Technology, funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant 13/RC/2106). We are grateful to ADAPT colleagues Anna Kostekidou, Julie Connelly, Clare Conran, among others, for their support to the project. We would like to thank Ms Sara Garvey and Ms Rachel Costello of Trinity College Dublin who prepared the English and French parts of corpora and helped improve the categorization. They also built an initial system using a nascent version of the corpora to help understand the performance that might be in both English and French. Special thanks to Dr Barbara Plank and Ms Shuying Lin for spotting the bug in the calculation of exact-match accuracy. Barbara also provided helpful comments on the use of these measures, among others, which are much appreciated. We are also grateful to Dr Heba Elfardy and her team from Amazon who identified several issues of the corpora and helped improve the corpora for the shared task. Finally, all participants are much appreciated to take part of this shared task and together help us better understand the task questions.

References

Barbara Plank. All-In-1 at IJCNLP-2017 Task 4: Short Text Classification with One Model for All Languages. In *the Proceedings of IJCNLP, Shared Tasks*, pp. 143–148, 2017. AFNLP.

Bentley, Michael and Batra, Soumya. Giving Voice to Office Customers: Best Practices in How Office Handles Verbatim Text Feedback. In *IEEE International Conference on Big Data*. pp. 3826–3832, 2016.

Burns, Michelle. (2016, February). Kampyle Introduces the NebulaCX Experience Optimizer. Retrieved from http://www.kampyle.com/kampyle-introduces-the-nebulacx-experience-optimizer/

Chao-Hong Liu, Declan Groves, Akira Hayakawa, Alberto Poncelas and Qun Liu. Understanding Meanings in Multilingual Customer Feedback. In *the Proceedings of First Workshop on Social Media and User Generated Content Machine Translation (Social MT 2017)*, Prague, Czech Republic.

Dushyanta Dhyani. OhioState at IJCNLP-2017 Task 4: Exploring Neural Architectures for Multilingual

Customer Feedback Analysis. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 170–173, 2017. AFNLP.

Equiniti. (2017, April). Complaints Management. Retrieved from https://www.equiniticharter.com/services/complaints-management/#.WOH5X2_yt0w

Freshdesk Inc. (2017, February). Creating and sending the Satisfaction Survey. Retrieved from https://support.freshdesk.com/support/solutions/articles/37886-creating-and-sending-the-satisfaction-survey

Heba Elfardy, Manisha Srivastava, Wei Xiao, Jared Kramer and Tarun Agarwal. Bingo at IJCNLP-2017 Task 4: Augmenting Data using Machine Translation for Cross-linguistic Customer Feedback Classification. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 59–66, 2017. AFNLP.

Inmoment. (2017, April). Software to Improve and Optimize the Customer Experience. Retrieved from http://www.inmoment.com/products/

Keatext Inc. (2016, September). Text Analytics Made Easy. Retrieved from http://www.keatext.ai/

Nan Wang, Jin Wang and Xuejie Zhang. YNU-HPCC at IJCNLP-2017 Task 4: Attention-based Bi-directional GRU Model for Customer Feedback Analysis Task of English. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 174–179, 2017. AFNLP.

Pintu Lohar, Koel Dutta Chowdhury, Haithem Afli, Mohammed Hasanuzzaman and Andy Way. ADAPT at IJCNLP-2017 Task 4: A Multinomial Naive Bayes Classification Approach for Customer Feedback Analysis task. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 161–169, 2017. AFNLP.

Potharaju, Rahul, Navendu Jain and Cristina Nita-Rotaru. Juggling the Jigsaw: Towards Automated Problem Inference from Network Trouble Tickets. In *10th USENIX Symposium on Network Systems Design and Implementation (NSDI 13)*. pp. 127–141, 2013.

Prathyusha Danda, Pruthwik Mishra, Silpa Kanneganti and Soujanya Lanka. IIIT-H at IJCNLP-2017 Task 4: Customer Feedback Analysis using Machine Learning and Neural Network Approaches. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 155–160, 2017. AFNLP.

Salameh, Mohammad, Saif M Mohammad, and Svetlana Kiritchenko. Sentiment after translation: A case-study on Arabic social media posts. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL,* pp. 767–777, 2015.

Shuying Lin, Huosheng Xie, Liang-Chih Yu and K. Robert Lai. SentiNLP at IJCNLP-2017 Task 4: Customer Feedback Analysis Using a Bi-LSTM-CNN Model. In *the Proceedings of IJCNLP, Shared Tasks,* pp. 149–154, 2017. AFNLP.

SurveyMonkey Inc. (2017, April). Customer Service and Satisfaction Survey. Retrieved from https://www.surveymonkey.com/r/BHM_Survey

UseResponse. (2017, April). Customer Service & Customer Support are best when automated. Retrieved from https://www.useresponse.com/

Yin, Dawei, Yuening Hu, Jiliang Tang, Tim Daly, Mianwei Zhou, Hua Ouyang, Jianhui Chen, Changsung Kang, Hongbo Deng, Chikashi Nobata, Jean-Mark Langlois, and Yi Chang. Ranking relevance in yahoo search. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 323–332, 2016. ACM.