ADAPT at IJCNLP-2017 Task 4: A Multinomial Naive Bayes Classification Approach for Customer Feedback Analysis task

Pintu Lohar, Koel Dutta Chowdhury, Haithem Afli, Mohammed Hasanuzzaman and Andy Way
ADAPT Centre
School of Computing
Dublin City University
Dublin, Ireland
{FirstName.LastName}@adaptcentre.ie

Abstract

In this age of the digital economy, promoting organisations attempt their best to engage the customers in the feedback provisioning process. With the assistance of customer insights, an organisation can develop a better product and provide a better service to its customer. In this paper, we analyse the real world samples of customer feedback from Microsoft Office customers in four languages, i.e., English, French, Spanish and Japanese and conclude a five-plus-one-classes categorisation (comment, request, bug, complaint, meaningless and undetermined) for meaning classification. The task is to determine what class(es) the customer feedback sentences should be annotated as in four languages. We propose following approaches to accomplish this task: (i) a multinomial naive bayes (MNB) approach for multilabel classification, (ii) MNB with one-vs-rest classifier approach, and (iii) the combination of the multilabel classification-based and the sentiment classification-based approach. Our best system produces F-scores of 0.67, 0.83, 0.72 and 0.7 for English, Spanish, French and Japanese, respectively. The results are competitive to the best ones for all languages and secure 3^{rd} and 5^{th} position for Japanese and French, respectively, among all submitted systems.

1 Introduction

With the rapid development of the Internet, the generation of online content has been near exponential during the last few years. A snapshot of the amount of online content generated per minute is shown in Figure 1. One of the items (shown by

Figure 1: Statistics of UGC generated per minute[2]

an arrow) in this figure shows that a huge amount of money ($\$750k$) is spent online per minute. Such an activity of the Internet users reflects how they are very much involved in online shopping. Due to this reason, industry sectors nowadays are more inclined to make use of online business development. For example, the big multinational companies (e.g., *Microsoft*[3], *Ebay*[4], *Amazon*[5] etc.) advertise and sell products via Internet. In response to this, customers frequently post product reviews on various websites in different languages. It is very important to understand the customers' behaviour because it provides marketers and business owners with insight that they can use to improve their business, products and overall cus-

[2] *Created by Lori Lewis, Vice President, Social Media - Cumulus Media/Westwood One*
[3] https://www.microsoft.com/
[4] https://www.ebay.com/
[5] https://www.amazon.com/

Proceedings of the 8th International Joint Conference on Natural Language Processing, Shared Tasks, pages 161–169,
Taipei, Taiwan, November 27 – December 1, 2017.

tomer experience. The present work is based on a joint ADAPT[6]-Microsoft research project, where the representative real world samples of customer feedback are extracted from Microsoft Office customers in four languages, i.e. English, French, Spanish and Japanese and concluded a five-plus-one-classes categorisation (comment, request, bug, complaint, meaningless and undetermined) for meaning classification. They prepared this corpus in order to provide an open resource for international customer feedback analysis. The task is to develop a system in order to find out which one among the provided six classes a customer feedback sentence belongs to. According to the criteria of classification, each feedback sentence must have at least one tag assigned to it. The sentence can also be annotated with multiple tags. We propose following three approaches to accomplish the task of feedback categorisation:

(i) the multinomial naive bayes (MNB) approach for multi-label classification,

(ii) the MNB with one-vs-rest classifier approach, and

(iii) the combination of the multilabel classification and the sentiment classification-based approach.

For sentiment classification, we use an automatic sentiment analysis tool (see Section 4.3). The experimental results show that the MNB with one-vs-rest classifier alone is sufficient enough to produce competitive results and hence becomes our best system among all the three approaches. Our system secures 3^{rd} and 5^{th} positions for Japanese and French, respectively.

The remainder of this paper is organised as follows. In Section 2, we provide a brief history of works in this field. Section 3 describe the process of customer feedback analysis along with some examples provided in this shared task. We provide a detailed description of the experiments in Section 4. The results are discussed in Section 5. Finally, we conclude and point out some possible future works in Section 6.

2 Related work

There is a number of research works done in the area of feedback analysis. For example, Bentley and Batra (2016) implement the Office Customer Voice system that combines classification, on-demand clustering and other machine learning techniques with a rich web user interface. They use this approach to solve the problem of finding the signal in the feedback posted by the Microsoft office users. The work in Potharaju et al. (2013) presents NetSieve, a problem inference system that aims to automatically analyse ticket text written in natural language to infer the problem symptoms, troubleshooting activities, and resolution actions. In Nasr et al. (2014), they contribute to the literature on Transformative Service Research and customer feedback management by studying the overlooked area of positive customer feedback impact on the well-being of service entities. Wu et al. (2015) perform following three steps for understanding the customer reviews: (i) collectively using multiple machine learning algorithms to pre-process review classification, (ii) selecting the reviews on which all machine learning algorithms cannot agree and assign them to humans to process, and (iii) the results from machine learning and crowd-sourcing are aggregated to be the final analysis results. The work in Morales-Ramirez et al. (2015) presents a user feedback ontology specified in ontoUML (Guizzardi (2005)). They focus on online feedback given by the users upon their experience in using a software service or application. Dalal and Zaveri (2014) propose an opinion mining system that can be used for both binary and fine-grained sentiment classifications of user reviews. Their technique extends the feature-based classification approach to incorporate the effect of various linguistic hedges by using fuzzy functions to emulate the effect of modifiers, concentrators, and dilators. The work presented in Hu and Liu (2004) aims at mining and summarising all the customer reviews of a product. They only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. Their task is performed in three steps: (i) mining product features that have been commented on by customers, (ii) identifying opinion sentences in each review and deciding whether each opinion sentence is positive or negative, and (iii) finally summarizing the results.

Customer feedback analysis has a strong interconnection with sentiment analysis as the feedback is essentially the customers' reactions to-

[6] https://www.adaptcentre.ie/

さらにインスタでの写真の縮小まで出来なくなりました	bug
設定しなおしても変わりません	complaint
早く直して下さい	request
4/28 相変わらず、たったひとりの人だけタグ付け出来ません	bug
編集で付けようとしても、どうしてもその人のだけ消えてしまう	bug, comment
La desinstalo definitivamente.	complaint
El taxista nos ha timado y ha hecho más del doble del trayecto.	complaint
Desde la última actualización, cuando voy a entrar nunca vuelve a cargar.	bug
La recomiendo para todos	comment
De resto, todo muy bueno.	comment

Figure 2: Some examples of Feedback sentences in Spanish and Japanese

wards the product they are using and hence conveys a specific sentiment (e.g., negative, neutral, positive etc.). The work in Fang and Zhan (2015) categorises sentiment polarity of the online product reviews collected from *Amazon.com* by performing the experiments for both sentence-level categorisation and review-level categorisation. Broß (2013) detect the individual product aspects reviewers have commented on and to decide whether the comments are rather positive or negative. They focus on the two main subtasks of aspect-oriented review mining: (i) identifying relevant product aspects, and (ii) determining and classifying expressions of sentiment. Gräbner et al. (2012) propose a system that performs the classification of customer reviews of hotels by means of a sentiment analysis. They elaborate on a process to extract a domain-specific lexicon of semantically relevant words based on a given corpus (Scharl et al. (2003); Pak and Paroubek (2010)). The resulting lexicon backs the sentiment analysis for generating a classification of the reviews.

3 Customer feedback analysis

Most app companies treat the contents of these reports as confidential materials and also regard things such as categorisation of customer feedback as business secrets. To the best of our knowledge, there are only few openly available categorisations from these app companies. One of them is the commonly used categorisation which could be found in many websites, i.e., the five-class Excellent-Good-Average-Fair-Poor (SurveyMonkey[7]). The other one is a combined categorisation of sentiment and responsiveness, i.e. another five-class Positive-Neutral-Negative-Answered-Unanswered, used by an app company Freshdesk[8]. There are many other categorisations for customer feedback analysis, however, most of them are not publicly available (e.g., Clarabridge[9], Inmoment[10])

To provide an open resource for international customer feedback analysis, the organisers of the shared task of customer feedback analysis in IJCNLP-2017 prepared a corpus using their proposed five-class categorisation of meanings as annotation scheme. As mentioned earlier in Section 1, a feedback sentence must have at least one tag and can also be annotated with multiple tags. Figure 2 shows some feedback examples in Spanish and Japanese provided by the organisers of the shared task. These examples are taken directly from the shared task webpage[11]. We can see from these examples that each sentence is most likely to be assigned one tag but it is also possible to assign more than one tag in case of multiple possibilities. For example, one of the sentences in Japanese in Figure 2 is assigned two tags (*bug* and *comment*).

4 Experiments

Statistics of the whole data sets in all four languages (i.e., English, Spanish, French and

[7] https://www.surveymonkey.com/r/BHM_Survey

[8] https://freshdesk.com/

[9] http://www.clarabridge.com/

[10] https://www.inmoment.com/

[11] https://sites.google.com/view/customer-feedback-analysis/

Japanese) is shown in Table 1. In this paper, we

Language	Train	Dev	Test
English	3,065	500	500
Spanish	1,631	301	299
French	1,950	400	400
English	1,526	250	300

Table 1: Data statistics per language

propose three different approaches (see Section 4.1, 4.2 and 4.3) to analyse the customer feedback in English. For the other languages, we apply the method that produces the best output for English feedback. In addition to this, we also apply this method to the available translations of the Spanish, French and Japanese feedback into English (see Section 4.4).

4.1 MNB classification

MNB is a specific instance of a Naive Bayes classifier which uses a multinomial distribution[12] for each of the features instead of referring to conditional independence of each of the features in the model. In this classification method, the distribution is estimated by considering the generative Naive Bayes principle, which assumes that the features are multinomially distributed in order to compute the probability of the document for each label and keep the label maximizing probability. Assuming the feature probabilities $P(x_i|c_j)$ are independent given the class c and for a document d represented as features x_1, x_2, ..., x_n; the equation for MNB can be written as follows:

$$C_{NB} = \arg\max_{c \in C} P(c_j) \prod_{x \in X} P(x|c) \quad (1)$$

Applying MNB classifiers to text classification, the equation can be represented as:

$$C_{NB} = \arg\max_{c_j \in C} P(c_j) \prod_{i \in positions} P(x_i|c_j) \quad (2)$$

where, *positions* ← *all word positions in test document*

For this task, we initially applied MNB classification method to label the whole training dataset in a single step. Subsequently, we also performed iterative process which is discussed in detail in the following section.

[12] https://web.stanford.edu/class/cs124/lec/naivebayes.pdf

4.2 MNB with one-vs-rest approach

This approach is a variation of MNB classification that works as follows:

(i) We apply an iterative process of one-vs-rest classification method. As there are total of six different categories available (*comment*, *complaint*, *meaningless*, *request*, *bug* and *undetermined*), we select one of these categories as *one*-class and consider the remaining as the *rest*-class. For example, we may select *comment* as *one*-class and treat the remaining as *rest*-class. These two groups of feedback can be considered as *comment* and *non-comment* classes, respectively. We then perform the classification process between the *comment* and the *non-comment* categories.

(ii) Once the feedback sentences are labeled as *comment* and *non-comment* classes, we opt out the sentences tagged as *comment* class and consider the rest with two new group of classes (for example, *complaint* and *non-complaint*). This iterative process continues until all the feedback sentences are assigned tags.

Figure 3 shows the iterative MNB classification using one-vs-rest approach. In each step, we classify the sentences into two categories exactly in the same way as discussed in step (i) and step (ii) above. We group the feedback into two classes; namely *comment* and *non-comment*. The *non-comment* class consists of other five remaining categories; (i) *complaint*, (ii) *meaningless*, (iii) *request*, (iv) *bug*, and (v) *undetermined*. Those sentences which are assigned *comment* tag are opted out and the remaining are considered for the next iteration. The process continued until only two categories are left (*bug* and *undetermined*) for classification. However, we can also begin with any other feedback categories instead of *comment*. The reason behind selecting *comment* is that in the initial experiments, our system performed better in tagging the *comment* class as compared to the other ones. It is easier to tag the feedback sentences that belong to this class because the total number of *comment* feedback is much more than that of any other classes and hence the system learns a better model for this class. The same is true for the other classes (i.e., *complaint*, *meaning-*

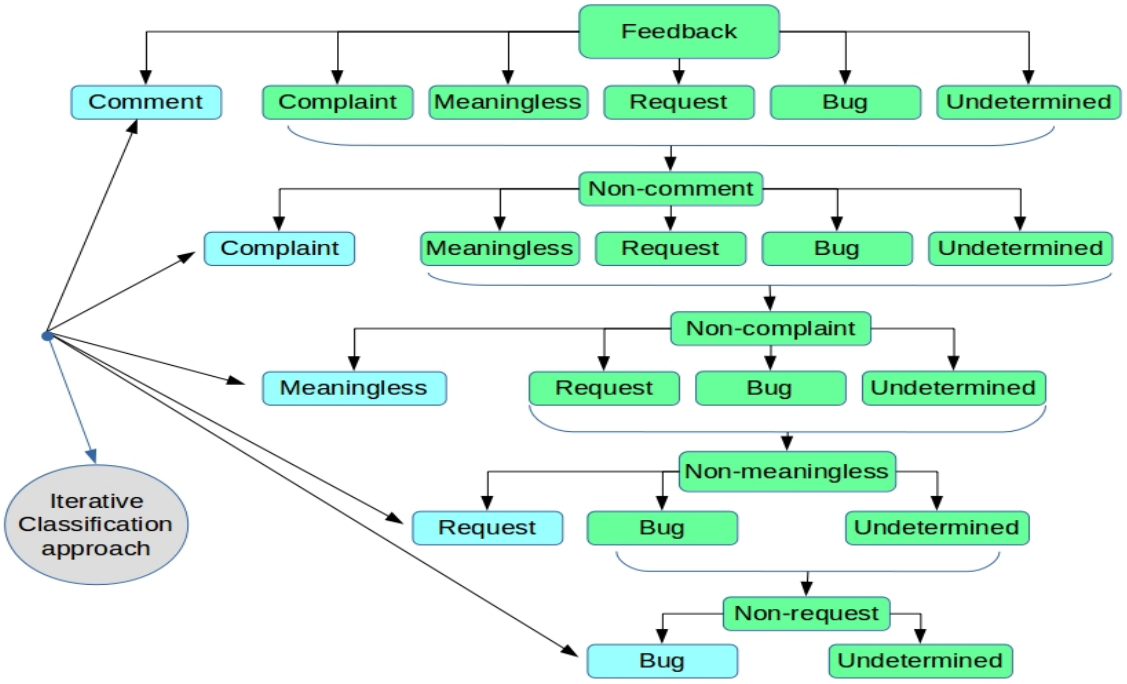

Figure 3: Iterative MNB classification with one-vs-rest approach

less, *request*, *bug* and *undetermined*) in the subsequent steps.

4.3 Multilabel classifier with sentiment classification

In addition to the approaches discussed in Section 4.1 and Section 4.2, we also apply sentiment classification approach and incorporate with the multilabel classification approach. We extract the sentiment scores (between 0 and 1, both inclusive) of all the feedback sentences using an automatic sentiment analysis tool (Afli et al., 2017), with 0 being extremely negative and 1 being extremely positive whereas any score close to 0.5 is considered to be neutral. Table 2 shows the observed sentiment range for different categories.

category	sentiment range
bug, complaint	0.2 to 0.6
meaningless, request	0.4 to 0.7
comment	0.4 to 0.9
undetermined	not fixed

Table 2: Sentiment range of Feedback categories

It is observed that some of the categories belong to overlapping ranges of sentiment scores. This observation implies that it is very difficult to identify a specific feedback category solely based on the sentiment scores due to the overlapping range of sentiment scores. However, it is visible in Table 2 that the *bug* and the *complaint* classes fall under the lower sentiment-score category. In contrast, *meaningless*, *request* and *comment* categories usually have higher sentiment scores, whereas the sentiment score for the *undetermined* category does not have any fixed range. Figure 4 shows the combination of multilabel classification and sentiment classification-based approach. Initially we filter out the *undetermined* and *meaningless* categories using the multilabel classification approach. The reasons behind performing this filtering are as follows: (i) the *undetermined* class has no fixed sentiment range, and (ii) the *meaningless* class has a specific sentiment range, but they are not related to customer feedback. This method works in following steps:

(i) Out of the six categories, *meaningless* and *undetermined* are filtered out using multilabel classification approach.

(ii) Sentiment scores are extracted for all the remaining feedback sentences.

(iii) Depending upon the sentiment scores, these feedback sentences are grouped into two different classes. For instance, if $score \geq 0.5$, the sentence is considered to be either *comment* or *request* class, and grouped together as *Comment_Request* category. In contrast,

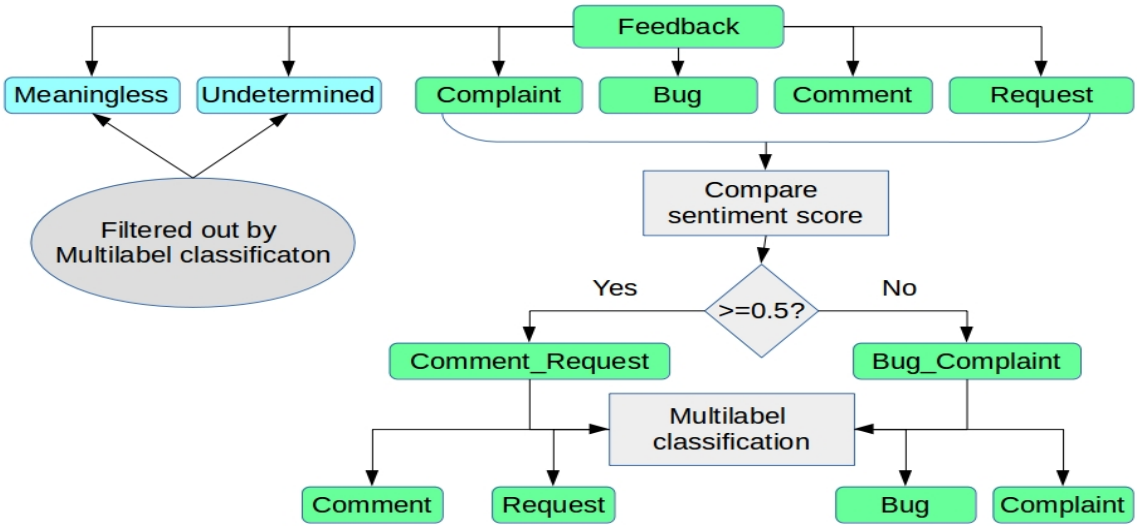

Figure 4: Machine learning combined with sentiment classification

if $score < 0.5$, it is treated as either *bug* or *complaint* class, and grouped together as *Bug_Complaint* category.

(iv) Finally, we apply multilabel classification to the *Comment_Request* category and identify each of *Comment* and *Request* classes. In a similar manner, each of *Bug* and *Complaint* classes are identified from the *Bug_Comment* group.

4.4 MNB for translated feedback

The translations of non-English (i.e., Spanish, French and Japanese) test data into English are also provided by the organisers. Since the translations of training data in these languages are not available, we train our classifier on English training data and then test it on the translations of Spanish, French and Japanese test data, respectively. Although our objective was to utilise the provided translations, this approach produces lower scores than other two approaches. The probable reason is that the trained model and the test data were originally in different languages.

5 Results

As the training and development data in English are larger than that of the other languages, we perform a series of different approaches (discussed in Section 4.1, Section 4.2 and Section 4.3) on English training data and applied the best approach to identify the feedback categories in other languages. These three approaches are termed as (i) MNB_all (for the MNB classification), (ii) ML_SentClass (for the multilabel classifier with sentiment classification approach), and (iii) MNB_one-vs-rest (for the MNB with one-vs-rest approach). It can be seen from

Systems	Precision	Recall	F1 score
MNB_all	0.674	0.6493	0.6614
ML_SentClass	0.6687	0.6455	0.6569
MNB_one_vs_rest	**0.692**	**0.6667**	**0.6791**

Table 3: Three different methods for English data

the table that the best results are obtained by the "MNB_one_vs_rest" approach. For the other two approaches the scores obtained are relatively lower due to the fact that (i) in *MNB_all* approach (see Section 4.1 for details), we identify the tags for all categories in a single step which is relatively more difficult task as compared to the iterative approach, and (ii) in *ML_SentClass* system (see Section 4.3), it is difficult to distinguish between the feedback categories based on their sentiment scores as some of them have overlapping range of scores as shown in Table 2. Based on the above observations, we apply the "MNB_one_vs_rest" approach to the other languages. However, each of the above three systems performs more or less similar in predicting the feedback sentences per category. For more detailed analysis, we provide Table 4 that highlights the performance of one of our systems per feedback category.

It can be observed for Table 4 that our system produce the best results for the *comment* class. Two probable reasons for this can be as follows: (i) all the feedback sentences belonging

Tag	Oracle count	Predicted	Correct	Precision	Recall	F1 score
comment	285	261	214	0.8199	0.7508	0.7838
complaint	145	178	103	0.5786	0.7103	0.6377
bug	10	11	2	0.1818	0.2	0.1904
meaningless	62	27	11	0.4074	0.1774	0.2471
request	13	23	7	0.3043	0.5384	0.3888
undetermined	4	0	0	0	0	0

Table 4: System performance per feedback category using MNB_one_vs_rest approach

to the *comment* category contain positive words, and (ii) it is seen that the majority of the training data belongs to the *comment* class. It therefore becomes easier for the system to learn the model that can better identify a feedback under this category as compared to the other ones. In contrast, the other categories share much smaller portion of the whole training data and hence it becomes more difficult to correctly identify these tags by the models learned from these categories. However, our system fails to identify any tag under *undetermined* category as the training data for this category is too small to train a classifier model.

Table 5 provides the summary of the results for English feedback. A total of 53 systems was submitted and this table shows some of them. The ranking was performed in the decreasing order of F1 scores. A total of 12 teams participated in this shared task in 4 languages mentioned above. The team names are published as coded names starting from "TA" to "TJ" and end with language code name (e.g., "EN" for English). The full name of a submission is represented as "<TX>-<method_name>-<land_ID>" where X can be any letter from "A" to "J". All of our submitted systems start with "TK". For example, our best performing system (highlighted in bold letters in Table 5) is named as "TK-MNB_one_vs_rest-EN" where "'MNB_one_vs_rest" is the method name and "EN" is the language ID (English in this case). Most of the teams submitted multiple runs for all the 4 languages. As mentioned earlier, we conduct 3 different experiments and applied the best one to the other language data sets. In addition to this, we also test our system on the translations of Spanish, French and Japanese feedback sentences into English. Therefore, we submitted 3 systems for English and 2 systems for each of the other languages but as mentioned earlier in Section 4.4, the models learned from the English data produce lower score when tested on the translated non-English test data into English. We therefore apply "MNB_one_vs_rest" method to the other languages. The performance of all the systems are evaluated using the precision, recall and F1 scores.

We can observe in Table 5 that out of the 53 submissions in English, our system (TK-MNB-one_vs_rest-EN) secures 18^{th} position with 0.6791 of $F1$ score whereas the highest $F1$ score achieved is 0.7557 and the lowest is 0.4175. Table 6, 7 and 8 show the results for Spanish, French and Japanese, respectively. For French and Spanish test data, our system achieves 5^{th} and 7^{th} rank with the F1 scores of 0.8361 and 0.7268, respectively. The system performs best for the Japanese data in terms of ranking and secures the 3^{rd} position. In overall, we can observe from the Table 5,6,7 and 8 that all the scores produced by our approach are relatively much closer to the highest scores than the lowest scores for all languages.

Rank	Systems	Precision	Recall	F1 score
1	TL-biLSTMCNN-EN	0.7485	0.7630	0.7557
2	TL-biCNN-EN	0.7383	0.7611	0.7495
...	...	...	...	...
18	**TK-MNB_one_vs_rest-EN**	0.692	0.6667	0.6791
19	TG-biLSTM-EN	0.6782	0.6782	0.6782
...	...	...	...	...
52	TB-M2-EN	0.4277	0.4277	0.4277
53	TB-M3-EN	0.4132	0.422	0.4175

Table 5: Results for English feedback sentences

Rank	Systems	Precision	Recall	F1 score
1	TA-M2-ES	0.8862	0.8862	0.8862
2	TA-M1-ES	0.8829	0.8829	0.8829
...	...	...	...	...
7	**TK-MNB_one_vs_rest-ES**	0.8361	0.8361	0.8361
8	TH-fastText-ES	0.8294	0.8294	0.8294
...	...	...	...	...
23	TF-NB-ES	0.5719	0.5719	0.5719
24	TF-SVM-ES	0.5719	0.5719	0.5719

Table 6: Results for Spanish feedback sentences

Finally, Table 9 highlights the summary of the overall results in all languages. It provides the

Rank	Systems	Precision	Recall	F1 score
1	TA-M1-FR	0.785	0.7476	0.7658
2	TC-CNN-FR-entrans	0.765	0.7286	0.7463
...	...	...	...	
5	**TK-MNB_one_vs_rest-FR**	0.745	0.7095	0.7268
6	TC-CNN-FR	0.735	0.7	0.7171
...	...	...	...	...
26	TF-NN-FR	0.515	0.4905	0.5024
27	TF-ss_predtest-FR	0.4875	0.4643	0.4756

Table 7: Results for French feedback sentences

Rank	Systems	Precision	Recall	F1 score
1	TA-M2-JP	0.7912	0.7507	0.7704
2	TA-M1-JP	0.777	0.7348	0.7553
3	**TK-MNB_one_vs_rest-JP**	0.7167	0.6869	0.7015
...	...	...	...	...
10	TK-ML_Trans-JP	0.6224	0.5847	0.603
11	TH-CNN-JP	0.58	0.5559	0.5677
...	...	...	...	...
23	TF-LR-JP-entrans	0.5367	0.5144	0.5253
24	TF-LR-JP	0.3267	0.3131	0.3197

Table 8: Results for Japanese feedback sentences

Lang	No. of submission	Our rank	Highest score	Baseline score	Lowest score	Our score
JP	24	**3**	0.7704	0.5933	0.3197	0.7015
FR	27	**5**	0.7658	0.6026	0.4756	0.7268
ES	24	7	0.8862	0.7726	0.5719	0.8361
EN	53	18	0.7557	0.5393	0.4175	0.6791

Table 9: Summary of results for all languages

comparison among the highest score, the baseline scores (provided by the organisers), the lowest scores and the scores produced by our system. For all languages, our system performs better than the baseline and produces comparable results to the top ones. Our system secures 3^{rd} and 5^{th} rank for Japanese and French, respectively. In overall, the scores produced by our approach are competitive and relatively much closer to the highest scores than the lowest scores for all languages.

6 Conclusions and Future work

In this work, we presented different approaches based on (i) multinomial naive bayes algorithm, and (ii) a combination of multilabel classification and sentiment analysis technique to identify the tags of a collection of customer feedback in four languages. Initially we applied three different approaches for customer feedback-classification in English. We then selected one of these approaches that produced the highest scores and applied it to the other languages. In addition to this, we also tested our system on the translations of the feedback sentences (non-English feedback into English). Our system produced competitive results for all of the languages and secured 3^{rd} and 5^{th} rank for Japanese and French, respectively in terms of F1 score. However, all of the methodologies were not tested on non-English datasets. In future, we plan to apply them to the other languages in order to see the effects in the results. We will also extend our study on the frameworks using other efficient machine learning techniques to develop a new algorithm utilising the benefits of the sentiment analyser so that it can be effectively used in prediction.

Acknowledgments

This research is supported by Science Foundation Ireland in the ADAPT Centre (Grant 13/RC/2106) (www.adaptcentre.ie) at Dublin City University.

References

Haithem Afli, Sorcha McGuire, and Andy Way. 2017. Sentiment translation for low resourced languages: Experiments on irish general election tweets. In *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*, Budapest, Hungary.

Michael Bentley and Soumya Batra. 2016. Giving voice to office customers: Best practices in how office handles verbatim text feedback. In *The fourth IEEE International Conference on Big Data*, pages 3826–3832, Washington DC, USA.

Jürgen Broß. 2013. *Aspect-Oriented Sentiment Analysis of Customer Reviews Using Distant Supervision Techniques*. Ph.D. thesis, Freie Universität Berlin, Berlin, Germany.

Mita K. Dalal and Mukesh A. Zaveri. 2014. Opinion mining from online user reviews using fuzzy linguistic hedges. *Applied Computational Intelligence and Soft Computing*, 2014:735942:1–735942:9.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5.

Dietmar Gräbner, Markus Zanker, Günther Fliedl, and Matthias Fuchs. 2012. Classification of customer reviews based on sentiment analysis. In *Proceedings of the International Conference on Information and Communication Technologies*, pages 460–470, Helsingborg, Sweden. Springer Vienna.

Giancarlo Guizzardi. 2005. *Ontological foundations for structural conceptual models*. Ph.D. thesis.

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, Seattle, Washington, USA. ACM.

Itzel Morales-Ramirez, Anna Perini, and Renata S. S. Guizzardi. 2015. An ontology of online user feedback in software engineering. *Applied Ontology*, 10:297–330.

Linda Nasr, Jamie Burton, Thorsten Gruber, and Jan Kitshoff. 2014. Exploring the impact of customer feedback on the well-being of service entities: A tsr perspective. *Journal of Service Management*, 25(4):531–555.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation*, pages 1320–1326, Valletta, Malta.

Rahul Potharaju, Navendu Jain, and Cristina Nita-Rotaru. 2013. Juggling the jigsaw: Towards automated problem inference from network trouble tickets. In *The 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 127–141, Lombard, Illinois, USA.

Arno Scharl, Irene Pollach, and Christian Bauer. 2003. Determining the semantic orientation of web-based corpora. In *4th International Conference on Intelligent Data Engineering and Automated Learning*, Hong Kong, China.

Heting Wu, Hailong Sun, Yili Fang, Kefan Hu, Yongqing Xie, Yangqiu Song, and Xudong Liu. 2015. Combining machine learning and crowdsourcing for better understanding commodity reviews. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 4220–4221.