

A corpus for aspect-based sentiment analysis in Vietnamese

1st Minh-Hao Nguyen

Faculty of Computer Science

University of Information Technology-VNUHCM

Ho Chi Minh, Vietnam

Email: haonm.12@grad.uit.edu.vn

2nd Tri Minh Nguyen

Faculty of Computer Science

University of Information Technology-VNUHCM

Ho Chi Minh, Vietnam

Email: 15520926@gm.uit.edu.vn

3rd Dang Van Thin

Multimedia Communications Laboratory (MMLab)

University of Information Technology-VNUHCM

Ho Chi Minh, Vietnam

Email: thindv@uit.edu.vn

4th Ngan Luu-Thuy Nguyen

Faculty of Computer Science

University of Information Technology-VNUHCM

Ho Chi Minh, Vietnam

Email: ngannlt@uit.edu.vn

Abstract—Recently, researchers have shown an increased interest in the aspect-based sentiment analysis problem. The goal is to extract valuable information concerning the aspects mentioned in users comments. This problem can be divided into three sub-tasks: term extraction, aspect detection, and polarity detection. In this paper, we present a new annotated corpus for studies on the two sub-tasks: aspect detection and polarity detection. Our corpus includes 7,828 restaurant reviews at document-level. We also performed a supervised learning method with rich features, achieving the F1-score of 87.13% for the aspect detection and the F1-score of 59.20% for polarity detection. Our corpus is published for research purpose¹.

Index Terms—Vietnamese corpus, aspect-based sentiment analysis

I. INTRODUCTION

With the development of information and communication technology, people tend to use social networks and websites to consult comments or suggestions about a product or service. However, it is difficult to refer to these comments before deciding to buy a product or use a service when there are a large number of comments on various aspects of the product. Two cases may occur: people only read a few reviews and then make a decision so that choice may be wrong. Another case is that the reader spends much time to read reviews in order to decide whether or not to use the service. The problem is how we can automatically analyze the reviews to support the decision making of customers, which is also the goal of Aspect-based sentiment analysis (ABSA). ABSA aims at identifying emotional factors on assessments related to specific aspects of a specific domain such as restaurant, hotel, laptop, etc.

For example, given a restaurant review in Vietnamese “Đồ ăn ngon nhưng hơi ít. Chắc mình nghĩ giá cả đồ ăn rẻ nên thế”, traditional sentiment analysis task classifies this review

into three categories: (positive), negative (negative) or neutral (neutral). This does not provide enough information about the restaurant. ABSA helps us to analyze more deeply into each aspect [1]. For the above example, we can mine the following information: the aspect *QUALITY* receives a positive comment thanks to the clause “đồ ăn thì ngon” (“Food is delicious”); the aspect *STYLE&OPTIONS* receives a negative comment because of the word “hơi ít” (“few”); and the last aspect *PRICES* is positive thanks to the clause “giá đồ ăn thì rẻ” (“The price of food is cheap”). When we analyze the status of each aspect, we can discover information from user reviews automatically.

In recent years, the ABSA task has attracted the research community in the world as well as researchers in Vietnam. Many corpora and algorithms have been developed for many languages, including English, Chinese, etc. However, for the Vietnamese, there are not yet many annotated corpora for the research on this problem. Therefore, in this paper, we present a new labeled corpus for the restaurant domain. Besides, we also implement a baseline method for evaluating our corpus.

The rest of the paper is structured as follows. Section 2 is an overview of related work about corpora and methods. It covers national and international studies on the ABSA problem. Section 3 presents the process of building the annotation guidelines and corpus. Then, we introduce the achieved corpus in Section 4. Section 5 presents the evaluation of a baseline method on our corpus. Finally, we draw a conclusion and future work in Section 6.

II. RELATED WORK

In order to study this problem, many corpora built in different languages. For the English language, there are corpora for domains such as electricity [2]; books, kitchen equipment, and electronic products [3]. Sorgente *et al.* [4] introduces an Italian ABSA corpus. Steinberger *et al.* and [5] introduce datasets for the Czech language for restaurant domain. Shared-tasks

¹<https://sites.google.com/uit.edu.vn/uit-nlp/corpora-projects?authuser=0>

on aspect-based sentiment analysis have been organized for research groups around the world to propose methods to solve problems. The SemEval ABSA shared-task was organized for 3 consecutive years in 2014 [6], 2015 [7] and 2016 [8].

For Vietnamese, the first shared-task on aspect-based sentiment analysis was organized by Vietnamese Language and Speech Processing community in 2018 [9]. For this shared task, two Vietnamese benchmark corpora were provided for the restaurant domain and the hotel domain. The best submission on both two domains considered the task as multiple binary classification problem [10]. For aspect detection, they built a single binary classification for each aspect and treated the polarity detection as a multi-class classification problem. Thin *et al.* [11] implemented a deep Convolutional Neural Network architecture for aspect detection task and experimented on the two above datasets. Their method outperformed the best submission at VLSP shared-task 2018 with 80.40% and 69.25% in F1-score for the restaurant domain and the hotel domain, respectively. After that, Thuy *et al.* [12] also presented a supervised method for cross-language aspect detection. They translated an ABSA corpus from English to Vietnamese via Google Translate tool and combined with a manually annotated corpus in Vietnamese. Their Vietnamese dataset contains 3,796 sentences of restaurant reviews. With sentences translated from English, their corpus included 6472 sentences for both English and Vietnamese. Besides, they also proposed using word embedding as features and demonstrated its effectiveness for aspect detection problem.

Mai and Le [13] collected and annotated Vietnamese ABSA corpora for two tasks: opinion target extraction and sentiment polarity detection for the smartphone domain. They presented a multi-task model for the two tasks by using the sequence-labeling scheme associated with bidirectional recurrent neural networks (BRNN) and conditional random field (CRF). Their experimental results showed that the BRNN-CRF model outperformed the CRF method with handcraft features.

For the education domain, Nguyen *et al.* [14] presented a UIT-VSFC corpus which consists of 16,000 sentences. This corpus was annotated by the human on two sub-tasks: sentiment-based task and topic-based task with the inter-annotator agreements 91.20% and 71.07%, respectively. In addition, they experimented with the Maximum Entropy classifier and got promising results for two tasks. Nguyen *et al.* [15] conducted a series of experiments on four supervised learning methods on UIT-VSFC corpus. Their experimental results show that BiLSTM algorithm outperformed other methods with F1-score of 92.0% for sentiment classification and 89.6% for topic classification.

III. ANNOTATION GUIDELINES

Our annotation guidelines are built based on the VLSP shared-task guidelines 2018 [9] and the SemEval 2016 guidelines [8]. However, unlike the previous corpus, in this corpus, we designate the number of aspect labels of eight, and the number of labels of polarity is five as follow:

A. Aspect labels

We have considered the number of aspects and the types of aspect based on some of the experts' recommendations about this domain. Each aspect will be described by a short guideline to support annotators identify the aspect in the review.

- **LOCATION:** The users mention the explicit and implicit location of the restaurant - regarding the area, address, and surroundings of the restaurant. For example "*Vị trí nhà hàng nằm ở trung tâm*" ("*The location of the restaurant is in the center*")
- **PRICE:** the review express directly or indirectly about prices of meal or service, e.g.,. For example "*Thức ăn đắt quá*" ("*Food is too expensive*")
- **SERVICE:** The opinions focus on customer service, kitchens, check-in counters in terms of speed and service quality of the general restaurant, delivery service, delivery attitude, staff attitude, protection, shopkeeper, labour. For example "*Nhân viên nhiệt tình*" ("*The staff is friendly*")
- **QUALITY:** The review refers to the quality of general restaurants or dishes, drinks including flavour, taste, freshness, crispness, consistency, softness, porosity, overall quality of food/drink in general and in particular. For example "*Thức ăn ngon*" ("*The food is delicious*")
- **AMBIENCE:** The users describe the view, the space or the atmosphere of the restaurant, related to entertainment, decoration or style the restaurant. For example "*Nhà hàng có không gian ấm cúng*" ("*The restaurant has a warm space*")
- **STYLE&OPTIONS:** The review is related to the presentation, the size of the meal, the food options/menu or the variety and richness of the food (eg., new, creative, vegetarian food). For example "*Thức ăn của nhà hàng khá ngon*" ("*The restaurant's food is quite good*")
- **MISCELLANEOUS:** The review presents other restaurant issues such as discount codes, chopsticks, dishes, tables and chairs, fans, restaurant equipment (fans, air conditioners), etc.,. For example "*Nhà hàng thường giảm giá 20% vào cuối tuần*" ("*The restaurant often discount 20% on weekends*")

B. Polarity labels

For each aspect, we use 5 level sentiment labels to represent its polarity: *very positive*, *positive*, *neutral*, *negative* and *very negative*.

- **Very positive:** The aspects are presented at a very positive level. For example, with the text "*Thức ăn của nhà hàng thì ngon tuyệt vời*" ("*The restaurant's food is delicious.*").
- **Positive:** The aspects are shown in a positive way. For example, "*Nhân viên phục vụ khá tốt*" ("*The staff serves quite good*").
- **Neutral:** The aspects are mentioned but expressed in a normal state, not complimented nor blamed. For instance, "*Không gian thì bình thường*" ("*Space is normal*").
- **Negative:** The aspects are commented at the level of negative. For example, "*vị trí nhà hàng thì khó tìm*" ("*the location of restaurant is hard to find*").

TABLE I
THE INTERANNOTATOR AGREEMENTS ON OUR CORPUS

| Task | P_o | P_e | A_m |
|------------------|-------|-------|-------|
| Aspect Detection | 95.58 | 37.96 | 92.87 |
| Aspect Polarity | 97.29 | 75.91 | 88.75 |

TABLE II
THE INFORMATION OF OUR CORPUS AFTER PREPROCESSING

| The information | Train | Development | Test |
|-----------------|--------|-------------|--------|
| N.o reviews | 5479 | 776 | 1573 |
| N.o Tokens | 287660 | 81198 | 161804 |
| N.o vocab | 8982 | 3464 | 4852 |
| N.o aspects | 17227 | 2458 | 4968 |
| Average Length | 52.5 | 52.3 | 51.4 |

- **Very Negative:** The aspects are described in the extremely negative state. For example, "*Đồ ăn có rất ít sự lựa chọn*" ("*Food has very little choice*").

IV. CORPUS INFORMATION

To collect data, we crawl user's comments about the restaurant from Foody². For the purpose of building a real corpus, we only remove the comments that are not grammatical (English reviews, unsigned reviews, etc.) in Vietnamese or are not related to the restaurant domain. Then, three annotators (2 graduate and 1 undergraduate students) are asked to independently annotate a piece of the corpus based on the guidelines. In case of disagreement, the final decision has been selected by an expert on this topic. After each annotated process, we conduct to calculate the agreement of annotators and discuss the ambiguous cases. We use the Cohen's kappa coefficient [16] to measure the inter-annotator agreement:

$$A_m = \frac{P_o - P_e}{1 - P_e} \quad (1)$$

where A_m is the inter-annotator agreement, P_o is the observed agreement between annotators and P_e is expected agreement. Table I shows our inter-annotator agreements on two tasks. For aspect category and aspect polarity, we finally achieved the A_m is 92.87% and 88.75% corresponding to each task. These values are considered as a satisfactory agreement.

Our corpus consists of 7828 reviews and is divided into three datasets: training, development, testing at the rate of 7/1/2. Because these tasks are the multi-label classification problem, therefore it is necessary to split our corpus into three datasets correctly. To address it, we use a strategy³ which is proposed by Sechidis *et al.* [17]. Table II presents an overview of our corpus, while Table III shows the distribution of the aspects in each corpus in detail. It is important to notice that the size of our corpus is bigger to the size of the different corpus in Vietnamese [9] [12] with the same purpose.

²Foody is a community for people to search, evaluate, comment on places to eat and drink in Vietnam. Links: <https://www.foody.vn/>

³<https://github.com/trent-b/iterative-stratification>

V. BASELINE METHOD AND EXPERIMENTS

A. Baseline method

In this section, we present a supervised learning approach as a baseline in two tasks. The objective of this approach is to detect the aspect categories (e.g. *location*, *price*, *service*, *ambience*, *quality*, *style&option* and *miscellaneous*) and its polarity (e.g. *very positive*, *positive*, *neutral*, *negative*, *very negative*) from the review at the same time. To address it concurrently, we employ a multi-class classification model corresponding to each aspect inspired by Saeidi *et al.* [18]. The output of the classification model is vectorized as a one-hot vector. Each element in the vector represents the sentiment of that aspect, and the first element indicated that the aspect is not assigned for the review. Text pre-processing is one of the key components in the text classification framework. In this paper, we conduct basic processing steps on reviews as replacing the money value, hashtag, website site, lower reviews. Word segmentation is an important step in Vietnamese processing; in this paper, we use Python Vietnamese Core NLP Toolkit⁴ to segment reviews into words. To train a supervised model for each aspect, we use Linear Support Vector Machine⁵ in conjunction with the following features extracted from the review:

- **Ngram features:** the n-grams of words are extracted from processed reviews. In this case, we choose bi-gram, tri-gram, and four-gram from the reviews.
- **Part-of-speech information:** We extract the information of part-of-speech of noun, verb, adjective as features.
- **Word features:** Nouns, verbs, and adjectives are also extracted from the reviews a feature.
- **Category Similarity:** For each aspect, we create a seed of words by selecting top 10 nearest words to category name in Vietnamese (e.g. *service* = "*phục vụ*"). Then we consider the maximum category similarity as a feature by calculating the cosine similarity of each word in the review with words in the seed. For the word embedding, We crawler large users comments for restaurant domain and use the implementation by Řehůřek and Sojka [19]. Our pre-trained word embedding includes 3.2 million sentences with approximately 53 million tokens. Table V lists the representative words for each aspect.

Then, we use the TF-IDF weighting to convert the list of features of reviews into a vector for the classification model. As shown in Table III, this is an imbalance problem between aspects of sentiment. Therefore, we apply the SMOTE technique [20] to balance the number of samples between labels.

B. Experimental Results and Discussion

To evaluate the quality of the baseline method, we use precision, recall, and F1-score. Table VI presents the results of our baseline method on our corpus for aspect detection and aspect polarity task, respectively. In general, with this method,

⁴<https://github.com/trungtv/pyvi>

⁵<https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html>

TABLE III
THE DISTRIBUTION OF THE ASPECTS IN OUR THREE DATASETS

| Aspect | Very positive | | | Positive | | | Neutral | | | Negative | | | Very negative | | |
|---------------|---------------|------|------|----------|-----|------|---------|-----|------|----------|-----|------|---------------|-----|------|
| | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test | Train | Dev | Test |
| Location | 353 | 57 | 97 | 116 | 11 | 36 | 153 | 20 | 37 | 120 | 12 | 37 | 54 | 11 | 20 |
| Prices | 676 | 82 | 178 | 375 | 51 | 121 | 810 | 109 | 227 | 527 | 82 | 156 | 276 | 43 | 87 |
| Service | 1780 | 257 | 545 | 345 | 53 | 95 | 252 | 38 | 73 | 228 | 33 | 76 | 659 | 88 | 170 |
| Quality | 2582 | 347 | 727 | 991 | 144 | 284 | 751 | 114 | 205 | 151 | 31 | 41 | 382 | 48 | 112 |
| Ambience | 1730 | 254 | 529 | 378 | 65 | 96 | 204 | 20 | 53 | 164 | 23 | 51 | 91 | 12 | 31 |
| Style&Option | 1194 | 170 | 370 | 257 | 33 | 70 | 103 | 15 | 30 | 199 | 30 | 47 | 213 | 40 | 62 |
| Miscellaneous | 394 | 69 | 110 | 330 | 45 | 89 | 43 | 4 | 9 | 141 | 19 | 28 | 200 | 28 | 68 |
| SUM = | 8709 | 1236 | 2556 | 2792 | 402 | 791 | 2316 | 320 | 634 | 1530 | 230 | 436 | 1875 | 270 | 550 |

TABLE IV
THE DETAILED RESULTS OF ASPECTS (IN %)

| Task | Measure | Ambience | Location | Quality | Style&option | Prices | Miscellaneous | Service |
|------------------|-----------|-------------------|----------|---------|----------------|--------|------------------|---------|
| Aspect Detection | Precision | 93.59 | 96.19 | 90.68 | 81.28 | 95.60 | 84.38 | 95.32 |
| | Recall | 88.54 | 44.69 | 98.83 | 61.49 | 87.63 | 17.76 | 93.63 |
| | F1-score | 91.00 | 61.03 | 94.58 | 70.01 | 91.44 | 29.35 | 94.47 |
| | Overall | Precision = 92.02 | | | Recall = 82.73 | | F1-score = 87.13 | |
| Aspect Polarity | Precision | 72.56 | 67.62 | 61.23 | 59.13 | 60.94 | 46.88 | 60.15 |
| | Recall | 68.64 | 31.42 | 66.74 | 44.73 | 55.86 | 09.87 | 59.08 |
| | F1-score | 70.55 | 42.90 | 63.87 | 50.93 | 58.29 | 16.30 | 59.61 |
| | Overall | Precision = 62.52 | | | Recall = 56.21 | | F1-score = 59.20 | |

TABLE V
THE LIST OF TEN REPRESENTATIVE WORDS FOR EACH ASPECT.

| Location | Price |
|---------------------------------|------------------------------|
| "địa_chi" "vị_trí" "kas" | "giá_cả" "giá_thành" "giá" |
| "tụ_điểm" "kichiiiiii" | "ăn_giá" "tiềns" "tiqf" |
| "faifo_buffet" "lý_tưởng" | "còm_côi" "nhiệt_xôi" |
| "điểm_prù" "nguyễ_huệ" "ddáng" | "_hợp" "trung_tí" |
| Service | Quality |
| "phụ_vụ" "pv" "ẹgvà" | "chấtlượng" "chất_lượng" |
| "hệ" "nhiệtvtfnh" "rui_vẽ" | "tương_xứng" "toco_nam_đồng" |
| "tea_happy" "bán_hàng" | "lg" "tiền" |
| "vớivkhách" "a_chủ_pvu" | "duy_sashimi" "chất_lượng" |
| | "phn_quốc" "ròisúp" |
| Ambience | Style&Option |
| "ko_gian" "gian" "kgian" | "chọn_lựa" "chọn" "lựa_chon" |
| "khung_cảnh" "hông_gian" "view" | "lựa" "chọn_ý" "đa_dạng" |
| "decor" "k_gian" | "món_hoa" "chicker" |
| "âm_củng" "thoáng" | "lựa_chòn" "chọn_lực" |

TABLE VI
THE RESULTS OF THE BASELINE METHOD ON OUR CORPUS

| Task | Precision | Recall | F1-score |
|------------------|-----------|--------|----------|
| Aspect Detection | 92.02% | 82.73% | 87.13% |
| Aspect Polarity | 62.52% | 56.21% | 59.20% |

we had the high F1-score (87.13% on average) for aspect detection and acceptable F1-score (59.20% on average) for aspect polarity. The detailed results of the aspects are shown in Table IV.

As shown in the Table IV, the baseline method achieved

the very satisfactory results especially for *Ambience*, *Quality*, *Prices* and *Service* for aspect detection task. However, the results for aspect polarity task do not achieve the expected results because the baseline method is not able to extract the sentiment features for each aspect. This is one of the limitations of the baseline method.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented the new annotated corpus for aspect-based sentiment analysis in Vietnamese for the restaurant domain. The corpus includes 7828 reviews with the high annotated agreement for two sub-tasks - aspect detection and aspect polarity. We believe that this corpus can be used to study and experiment with deep neural architecture. On the other hand, we also experimented a baseline method with various features and achieved the F1-score 87.13% and 59.20% for aspect detection and aspect polarity, respectively. The results of the baseline method on the aspect detection task give a positive score; however, the results for the aspect polarity task is not as expected.

In future work, we will study the effectiveness of the deep neural network models on this corpus and public the corpus for language processing community with free research purpose. Besides, we also provide a more detailed analysis, guidelines annotation about this corpus.

ACKNOWLEDGMENT

This research is funded by the University of Information Technology - Vietnam National University Ho Chi Minh City under grant number B2019-26-01.

REFERENCES

- [1] M. Hu and B. Liu, "Mining and summarizing customer reviews", in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004, pp. 168–177.
- [2] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews", in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05, Vancouver, British Columbia, Canada: Association for Computational Linguistics, 2005, pp. 339–346.
- [3] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification", in *Proceedings of the 45th annual meeting of the association of computational linguistics*, 2007, pp. 440–447.
- [4] A. Sorgente, V. C. Flegrei, G. Vettigli, and F. Mele, "An italian corpus for aspect based sentiment analysis of movie reviews",
- [5] J. Steinberger, T. Brychcin, and M. Konkol, "Aspect-level sentiment analysis in czech", in *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2014, pp. 24–30.
- [6] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "Semeval-2014 task 4: Aspect based sentiment analysis", in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 27–35.
- [7] M. Pontiki, D. Galanis, H. Papageorgiou, S. Manandhar, and I. Androutsopoulos, "Semeval-2015 task 12: Aspect based sentiment analysis", in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 486–495.
- [8] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, A.-S. Mohammad, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, *et al.*, "Semeval-2016 task 5: Aspect based sentiment analysis", in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 19–30.
- [9] H. Nguyen, H. Nguyen, Q. Ngo, L. Vu, V. Tran, B. Ngo, and C. Le, "Vlsp shared task: Sentiment analysis", *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 295–310, 2019.
- [10] D. V. Thin, V. Nguyen, N. Kiet, and N. Ngan, "A transformation method for aspect-based sentiment analysis", *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 323–333, 2019.
- [11] D. V. Thin, V. D. Nguye, K. V. Nguyen, and N. L. Nguyen, "Deep learning for aspect detection on vietnamese reviews", in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, Nov. 2018, pp. 104–109.
- [12] N. T. T. Thuy, N. X. Bach, and T. M. Phuong, "Cross-language aspect extraction for opinion mining", in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Nov. 2018, pp. 67–72.
- [13] L. Mai and B. Le, "Aspect-Based Sentiment Analysis of Vietnamese Texts with Deep Learning", in *Intelligent Information and Database Systems*, N. T. Nguyen, D. H. Hoang, T.-P. Hong, H. Pham, and B. Trawiński, Eds., Cham: Springer International Publishing, 2018, pp. 149–158.
- [14] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L. Nguyen, "Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis", in *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, Nov. 2018, pp. 19–24.
- [15] P. X. V. Nguyen, T. T. T. Hong, K. V. Nguyen, and N. L. Nguyen, "Deep learning versus traditional classifiers on vietnamese students' feedback corpus", in *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, Nov. 2018, pp. 75–80.
- [16] P. K. Bhowmick, P. Mitra, and A. Basu, "An agreement measure for determining inter-annotator reliability of human judgements on affective text", in *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, Association for Computational Linguistics, 2008, pp. 58–65.
- [17] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the stratification of multi-label data", in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2011, pp. 145–158.
- [18] M. Saeidi, G. Bouchard, M. Liakata, and S. Riedel, "Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods", in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1546–1556.
- [19] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora", in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, ELRA, May 2010, pp. 45–50.
- [20] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun. 2002.