Using machine translation to improve sentiment classification task of polish restaurant reviews

Tomasz Starakiewicz
Faculty of Economic Sciences
University of Warsaw
tomasz.starakiewicz@student.uw.edu.pl

Dustin Pacholleck
Faculty of Economic Sciences
University of Warsaw
d.pacholleck@student.uw.edu.pl

***Abstract*—This electronic document is a "live" template and already defines the components of your paper [title, text, heads, etc.] in its style sheet.**

I. Introduction

xxx

II. Literature review

xxx

-
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

III. Data

A. Sourcing

The main data source we use in the paper is Google Maps. We accessed the Google Places API in October 2022 and performed an exhaustive search of all restaurants in Poland. The API was limited to returning a maximum of 60 restaurants, paged 20 results at a time, around a single set of coordinates submitted in the API call. To overcome this limitation, we first split the territory of Poland into hexagons using the H3 library created by Uber. Then we called the API at the center of each hexagon. Given that the results were sorted according to the center from the submitted coordinates, if any of the restaurants returned was outside of a circle circumscribed around the hexagon, the hexagon in question was considered exhausted. If the hexagon was not exhausted, we proceeded to recursively split it into 7 smaller hexagons and call the API for each of the resulting hexagon centers. An exemplary result of the procedure, limited to Warsaw is shown on Fig. 1.

Fig. 1. Exhausted hexagons in Warsaw

The resulting universe contained 41,817 restaurants, shown on Fig. 2.

For each of the restaurants, we accessed the review section at the Google Maps page and scraped it using Selenium. As a result, we were able to collect 6.1M of non-empty text reviews, which represent all reviews of all Polish restaurants present on Google Maps as of October 2022. We collected both the reviews in their original languages, as well as their English translation powered by Google Neural Machine Translation, which is based on transformer architecture.

For the analysis, we selected only reviews longer than 150 characters, which reduced the number of reviews to 827,341. We also used Compact Language Detector *cld3* created by Google to identify non-Polish reviews. Cld3 is a pre-trained neural network model which uses uni-, bi- and trigrams fed into an embedding layer, then single hidden layer and finally into a softmax layer. We included only reviews classified as "Polish" with probability above 0.5. This reduced the number of reviews to 792,971.

B. Sampling

Each of the text reviews is accompanied by a rating on one to five scale. As expected, the distribution of ratings was highly imbalanced, as shown on Fig. 3, with just two classes responsible for 73% of all the ratings. Interestingly, however, the ratings also displayed significant polarization, since these two classes corresponded to the extreme ends of the rating sclae. The reviews were downsampled, with 12,500 reviews drawn randomly from each class, ensuring balanced dataset with equal representation of each rating. This resulted in the final dataset of 62,500 reviews, with 50,000 earmarked for training dataset and 12,500 for the test dataset.

C. Preprocessing

Both original and translated reviews were processed using pre-trained pipelines provided along with SpaCy library. SpaCy is industrial-grade NLP library offering solutions to process text at scale, using best-in-class methods for each language task. We have chosen the library to better represent real-life use cases, since the library is widely used in practice in the industry.

Two pipelines used for preprocessing included a mix of neural network and rule-based components. The reviews were tokenized, then converted into vector representations with 300 dimensions, tagged for parts-of-speech, parsed, and finally fed into lemmatization component. For each of the languages we prepared three versions of tokenized reviews: raw (with punctuation removed), lemmatized, and lemmatized with stop-words removed.

IV. Modeling

To predict the sentiment, we have selected two models commonly used in sentiment analysis: Naïve Bayes (NB) classifier and eXtreme Gradient Boosting (XGBoost) classifier. NB classifier serves as a baseline, due to its popularity as a text classification tool thanks to its speed combined with reasonable effectiveness. We chose XGBoost as a comparator, as it allows for modelling complex, non-linear interactions between features, which is a natural fit for text classification task.

A. Naïve Bayes classifier

Xxx

For NB classification, tokenized review text was converted into a matrix of token counts, with dictionary based on training data and tokens occurring in fewer than 50 reviews ignored. We considered two candidate models: Multinomial

NB classifier and Binary Multinomial NB classifier. To train the latter, we clipped the token frequencies at document level to 1. This is based on a common assumption in sentiment analysis, that word occurrence is more important that word frequency. The standard Multinomial NB classifier outperformed the Binary NB however, so we present only the results of the former.

B. *XGBoost classifier*

For classification with XGBoost, we converted the tokenized text reviews into a term frequency-inverse document frequency matrix. We ignored tokens occurring in fewer than 150 documents to reduce computational power required to train the model. We then performed a hyperparameter optimization

C. *Figures and Tables*

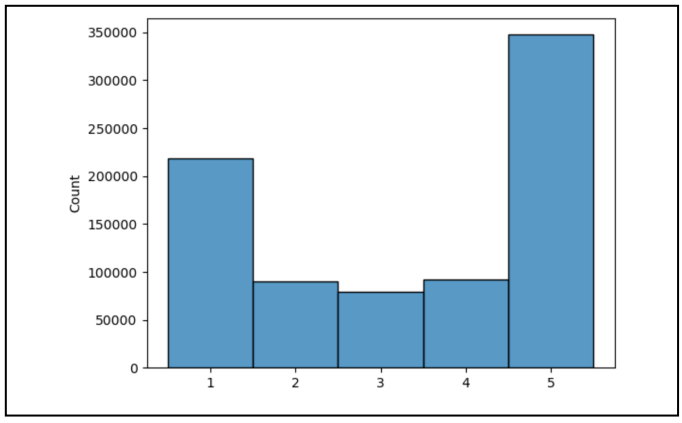
a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 1", even at the beginning of a sentence.

V. Results

TABLE I. NB-CLASSIFIER TEST RESULTS

Rating	**Test F1 scores for Naïve Bayes classifier**					
	Raw		***Lemmatized***		***Lemmatized with stopwords removed***	
(O)riginal (T)ranslated	**O**	**T**	**O**	**T**	**O**	**T**
1	0.65	0.64	0.64	0.63	0.63	0.61
2	0.43	0.42	0.43	0.42	0.41	0.40
3	0.43	0.41	0.41	0.40	0.41	0.40
4	0.57	0.55	0.56	0.54	0.55	0.53
5	0.80	0.78	0.79	0.78	0.77	0.76
Macro avg.	**0.58**	**0.56**	**0.57**	**0.55**	**0.55**	**0.54**

Fig. 2. Exhausted hexagons in Warsaw



[a] Sample of a Table footnote. (*Table footnote*)

Acknowledgment *(Heading 5)*

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

xxx

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.