Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/978-1-61499-442-8-126

Uses of Machine Translation in the Sentiment Analysis of Tweets

Jānis PEISENIEKS[a, 1] and Raivis SKADIŅŠ[b]
[a] *University of Latvia, Latvia*
[b] *Tilde, Latvia*

Abstract. This paper reports on the viability of using machine translation (MT) for determining the original sentiment of tweets, when translating tweets made in internationally less used language into more frequently used ones. The results of the study show that it is possible to use MT and sentiment analysis (SA) systems to produce SA results with significant precision.

Keywords. Machine translation, sentiment analysis, Latvian

Introduction

Sentiment analysis has been a very active field of study lately, and there have been a lot of high quality tools developed for English. These tools allow for near real-time analysis of various user generated data sources like online reviews, blogs, news, and social networks. By using sentiment analysis on these data sources, it is possible to gain a deeper understanding of social processes and help businesses and governments make well informed decisions. However the main problem of these tools is that they are usually created for large, internationally used languages.

We wanted to see whether it was possible to use publicly available MT and SA systems to discern the sentiment of a tweet that was written in an internationally less used language, in this case, Latvian.

1. Related Work

One of the main problems in cross-language sentiment translation is the quality of the translation software and whether translations obtained using MT can be used in sentiment analysis. To this aim, supervised Machine Translation systems have been used on the English language [1] to produce training data for other languages. On the other hand, a recent study [2] has shown, that the quality of such work can be sub-optimal.

Balahur, Turchi et. al. [3] have also shown that it is possible to use MT systems to train multilingual sentiment classifiers, where the training of classifiers in one language also improves the abilities of classifiers in other languages.

[1] Corresponding Author: Jānis Peisenieks, 29 Raina Blvd., Riga, Latvia; E-mail: janis@peisenieks.lv

2. Test Corpus

In order to study the viability of the proposed approach, a well annotated test corpus of tweets written in Latvian is necessary. The authors created such a corpus, because no such corpus was publicly available.

By using a custom crowd sourcing website, a tweet test corpus with sentiment polarity annotation was created with 3 sentiment classes: positive, neutral, and negative. The tweets for this website were gathered from Twitter's real-time API from November 2013 to March 2014, using a rough contour of Latvia as the query for the Twitter API. Furthermore, the Twitter API provides a language for each tweet, and only tweets with Latvian or no language were used.

In order to assess the sentiment of a tweet accurately, each tweet was rated 11 times (some tweets have up to 13 ratings due to real time specifics of the website), and only the ones with strong annotator consensus were included in the test corpus. In total, ~20,000 tweet ratings were processed.

These ~20,000 ratings produced 1,722 adequately rated tweets, and of these, 1,177 had the required annotator consensus to be included in the test corpus.

Table 1. Distribution of tweets in sentiment classes in the test corpus

Sentiment class	Tweet count
Positive	383
Neutral	627
Negative	167

The distribution of these tweets (Table 1) reflects the distribution of tweets being tweeted in Latvian. The whole test corpus has been made publicly available on the code collaboration website Github.com[2].

Additionally, Fleiss's kappa [4] was used to measure the reliability of agreement between the annotators and is 0.284, which, according to Landis and Koch [5], can only be described as fair agreement. This could have been influenced by the lack of the "skip" functionality in the crowd-sourcing website, which would allow the annotator to skip the current tweet if it contained no sentiment information.

3. Uses of MT in SA

To study the viability of the proposed approach, 3 publicly available MT and SA systems were chosen.

The MT systems used in this study were chosen based on their performance of LV-EN translations [6], and those are Google Translate, Bing Translator, and Tilde Translator. The whole of the test corpus was translated from Latvian to English using each of the MT systems.

In order to assess the sentiment of the translated tweets, 3 publicly available SA systems (AlchemyAPI, Textalytics, and Semantria) were chosen, based on their performance [7] and ease of use. The publicly available SDKs for each of the SA

[2] https://github.com/FnTm/latvian-tweet-sentiment-corpus

systems were used to analyze the sentiment of all of the translations, thus producing MT+SA system pairs, the performance of which could be evaluated.

4. Evaluation and Results

We used the standard performance metrics, such as precision, recall, and F_1-measure (Table 2), to evaluate the SA results. While processing the SA results, it became apparent that not all of the SA tools properly process the neutral sentiment class, which means that the SA tool would either produce results that are indistinguishable from no sentiment or that are hard to identify. Additionally, the SA systems that could at least partially recognize the neutral sentiment class had a very low precision and extremely low recall, which could be caused by the difficulty to distinguish text with a neutral sentiment from text with no sentiment. Because of this, all of the further research/interpretation was done only on positive and negative sentiment classes.

Table 2. Precision, recall, and F_1-measure of MT+SA system pairs

Sentiment class / MT+SA system pair	Positive			Negative			Neutral		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
Google Textalytics	0.64	0.70	0.67	0.62	0.39	0.48	0.20	0.00	0.01
Tilde Textalytics	0.64	0.61	0.63	0.58	0.39	0.47	0.12	0.00	0.01
Bing Textalytics	0.55	0.74	0.63	0.59	0.36	0.45	0.21	0.01	0.02
Google AlchemyAPI	0.46	0.74	0.57	0.38	0.75	0.50	-	-	-
Tilde AlchemyAPI	0.45	0.74	0.56	0.38	0.69	0.49	-	-	-
Bing AlchemyAPI	0.46	0.78	0.58	0.39	0.72	0.51	-	-	-
Google Semantria	0.63	0.59	0.61	0.49	0.40	0.44	0.35	0.02	0.03
Tilde Semantria	0.61	0.47	0.53	0.48	0.35	0.40	0.50	0.02	0.04
Bing Semantria	0.65	0.59	0.62	0.50	0.42	0.45	0.33	0.01	0.02

Additionally, the overall accuracy of the MT+SA pairs was measured (Table 3) as the total percentage of tweets classified correctly. For easier interpretation, this same data has been graphed in Figure 1.

Table 3. Overall Accuracy of MT+SA pairs

MT+SA system pair	Accuracy	Confidence interval ±
Tilde Textalytics	54.73%	4.16%
Tilde Semantria	45.64%	4.16%
Tilde AlchemyAPI	72.55%	3.73%
Google Textalytics	61.27%	4.07%
Google Semantria	55.45%	4.15%
Google AlchemyAPI	74.55%	3.64%
Bing Textalytics	63.27%	4.03%
Bing Semantria	54.73%	4.16%
Bing AlchemyAPI	76.00%	3.57%

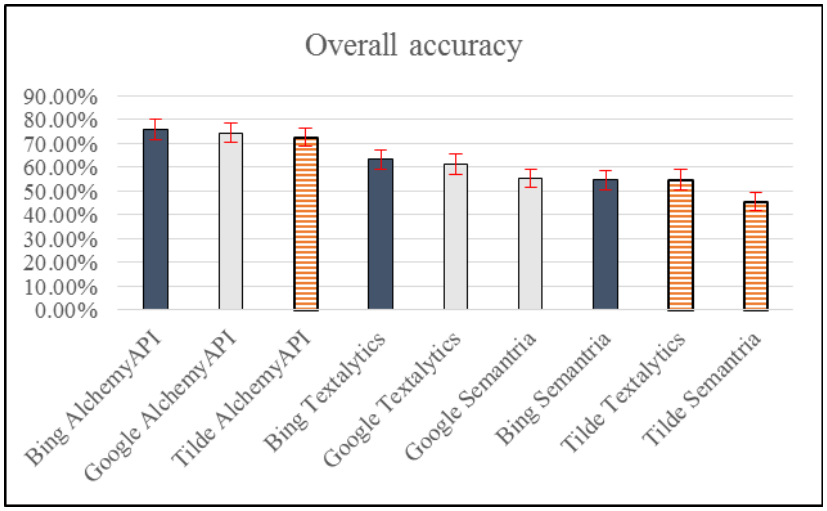

Figure 1. Overall Accuracy of MT+SA pairs

As can be seen from these results and from the standpoint of SA systems, AlchemyAPI is the clear leader, with the best SA results irrespective of the MT system used to translate the tweets. From the standpoint of MT systems, Bing Translator is the MT system that acquired the best results, however the results are not decisive in this matter.

Interestingly, the MT+SA pair with the highest overall accuracy (Bing Translator + AlchemyAPI) did not produce the best results in other metrics. One part of the answer to this question is that the mentioned pair is not the best, however it is consistent. However, it is possible that the way in which the experiments were conducted, accompanied with the characteristics of the test set, allowed for an interesting error to occur.

When looking at the SA systems, only AlchemyAPI does not even try to process the neutral sentiment class, which allows for a more broad classification in the other sentiment classes. This means that AlchemyAPI classifies more tweets as having either a positive or negative sentiment, where the other SA systems would produce no sentiment for the particular tweet. This can be easily seen in Figure 2 which shows the percentage of classified tweets that are false-positives. This means that a tweet that has been classified, for example, as a positive, is in fact from the neutral sentiment class.

Thus, even though AlchemyAPI provides for a higher overall accuracy, it also has more problems of correctly identifying neutral data. Also interestingly, the percentage of false positives between the positive and negative sentiment classes is quite small (~7%), which means that even though a lot of tweets have been incorrectly classified, the false-positives introduce very little bias to the data. This could mean that these false positives would have minimal impact in real-world use cases.

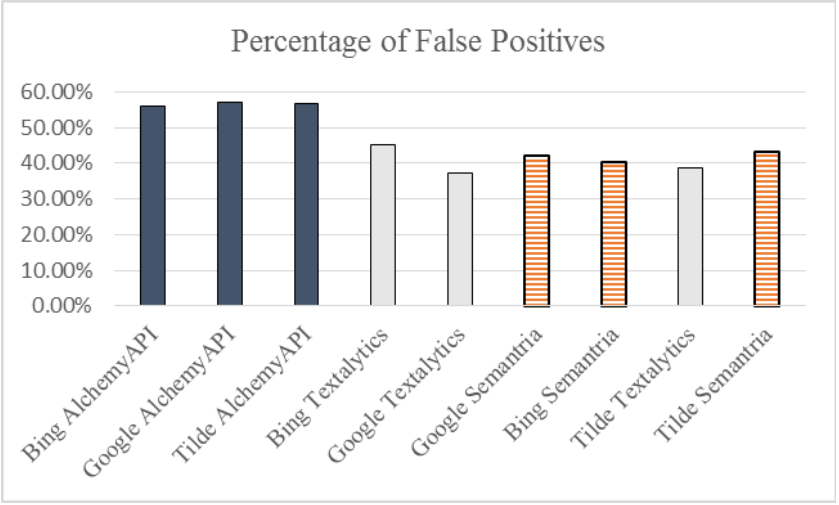

Figure 2. Percentage of false positives in classified tweets

5. Conclusions

The evaluation of the proposed approach shows that it is possible to do binary sentiment analysis on tweets originally written in a less internationally used language, with a high degree of accuracy. Additionally, it is clear that doing this sort of classification using 3 sentiment classes would provide results with a low degree of accuracy.

It should also be noted that during the binary sentiment analysis of tweets, a significant amount of false positives occur. Depending on the particular use case, this may or may not present problems and skew the end results.

References

[1] Balahur, A., and Turchi. M. (2012). Multilingual Sentiment Analysis using Machine Translation?. In Proceedings of the Association for Computational Linguistics: 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis Workshop (pp. 52-60) Jeju, Republic of Korea: ACL

[2] Cieliebak, M., Dürr, O., and Uzdilli. F. (2013) Potential and Limitations of Commercial Sentiment Detection Tools. Proceedings of the First International Workshop on Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM 2013)

[3] Balahur, A., Turchi, M., Steinberger, R., Perea-Ortega, J. M., Jacquet, G., Küçük, D., Zavarella, V., & El Ghali, A. (2014). Resource Creation and Evaluation for Multilingual Sentiment Analysis in Social Media Texts. In Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland: European Language Resources Association (ELRA)

[4] Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin. 76:378-382.

[5] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics. 1977 Mar; 33(1):159-74.

[6] Skadiņš, R., Šics, V., Rozis, R. (2014). Building the world's best general-domain MT for Baltic languages. In Human Language Technologies – The Baltic Perspective - Proceedings of the Sixth International Conference Baltic HLT 2014. Kaunas, Lithuania: IOS Press.
[7] Abbasi, A., Ammar, H., and Dhar, M. (2014). Benchmarking Twitter Sentiment Analysis Tools. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (pp. 823-829). Reykjavik, Iceland: European Language Resources Association (ELRA)