An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis

Matheus Araújo
Federal University of Minas Gerais
Belo Horizonte, Brazil
matheus.araujo@dcc.ufmg.br

Julio C. S. Reis
Federal University of Minas Gerais
Belo Horizonte, Brazil
julio.reis@dcc.ufmg.br

Adriano C. M. Pereira
Federal University of Minas Gerais
Belo Horizonte, Brazil
adrianoc@dcc.ufmg.br

Fabrício Benevenuto
Federal University of Minas Gerais
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Sentiment analysis has become a key tool for several social media applications, including analysis of user's opinions about products and services, support to politics during campaigns and even for market trending. There are multiple existing sentiment analysis methods that explore different techniques, usually relying on lexical resources or learning approaches. Despite the large interest on this theme and amount of research efforts in the field, almost all existing methods are designed to work with only English content. Most existing strategies in specific languages consist of adapting existing lexical resources, without presenting proper validations and basic baseline comparisons. In this paper, we take a different step into this field. We focus on evaluating existing efforts proposed to do language specific sentiment analysis. To do it, we evaluated twenty-one methods for sentence-level sentiment analysis proposed for English, comparing them with two language-specific methods. Based on nine language-specific datasets, we provide an extensive quantitative analysis of existing multi-language approaches. Our main result suggests that simply translating the input text on a specific language to English and then using one of the existing English methods can be better than the existing language specific efforts evaluated. We also rank those implementations comparing their prediction performance and identifying the methods that acquired the best results using machine translation across different languages. As a final contribution to the research community, we release our codes and datasets. We hope our effort can help sentiment analysis to become English independent.

Keywords

Sentiment Analysis; Multiple Languages; Machine Translation

1. INTRODUCTION

SAC 2016,April 04-08, 2016, Pisa, Italy

http://dx.doi.org/10.1145/2851613.2851817

Online Social Networks (OSNs) have been used by billion of users worldwide, being the most popular Web application nowadays [25]. On those systems, users can discuss an enormous variety of subjects, expressing their opinions, politic view, and even some subjective concepts. Because of the massive popularity and quantity of data shared on those systems, a variety of applications have emerged, aiming at extracting opinions and inferring public sentiments.

In this context, sentiment analysis has become a popular tool for data analysts, specially those that deal with social media data. It is common to find public opinion and reviews of services, events, and brands on social media. From the extracted data, sentiment analysis techniques can infer how people feel about a specific target, which is essential for companies aiming at focusing their investments on incorporating those potential clients and creating a more specific but also massive public marketing. Thus, sentiment analysis became a hot topic in Web applications, with the high demand from industry and academy, motivating the proposal of new methods to deal with this subject.

Despite the large interest from industry and academy, some substantial effort has been focused on sentiment analysis solutions that are English dependable, since this language is dominant in the Web content [31]. However, the potential market for sentiment analysis in different languages is vast. For example, suppose a mobile application that simply uses sentiment analysis. To leverage the application to multiple languages and several countries, one would require dealing with sentiment analysis approaches on multiple languages as well, which is currently quite limited. Some efforts even attempt to develop techniques to analyse sentiments from other specific languages: Arabic [3], German [29], Portuguese [33], Russian [39], among others. However, little is known about the performance prediction, viability and real need of those methods. More important, a different solution on each specific language is unfeasible for those interested simple in using sentiment as part of a system or application developed in multiple languages.

In this work, we investigate how a simple strategy can address the problem of sentiment analysis in multiple languages. Particularly, we analyse how the use of machine translation systems - such as Google Translate[1] - can affect the performance of English Sen-

[1]https://translate.google.com

timent Analysis methods in non-English datasets. Recent efforts show that Google Translate has a good performance to European languages but Asian languages are relatively poor [4]. We evaluate the prediction performance of twenty one sentiment analysis methods recently evaluated in a benchmark study [30] - AFINN, Combined, Emoticon Distant Supervisor, Emolex, Emoticons, Happiness Index, LIWC, NRC Hashtag, OpinionFinder, OpinionLexicon, Panas-t, Pattern.en, SANN, SASA, SenticNet, Sentiment140 Lexicon, SentiStrength, SO-CAL, Stanford Recursive Deep Model, SentiWordNet, Umigon and Vader - across nine different languages: Portuguese, French, Spanish, Italian, Turkish, Russian, Arabic, Dutch and German. According to *Internet World Stats*[2], six of those languages appear among the top ten languages used on the Web and represent 70% of the non-English content.

Despite the still large existent space for improvement in current state-of-the-art sentiment analysis methods for English, as suggested by a recent benchmark study [30], our findings suggest that machine translation systems are mature enough to produce reliably translations to English that can be used for sentence-level sentiment analysis and obtain lower, but still competitive prediction performance results. Additionally, we show that some popular language-specific methods do not have significant advantage over a machine translation approach. Our results also identify, across multiple language-specific datasets, the most suitable sentiment analysis methods designed English that perform well with machine translation.

The rest of the paper is organized as follows. Section 2 describes related work. In Section 3 we briefly present our methodology and Section 4 covers results. Finally, Section 5 concludes the paper.

2. RELATED WORK

Most approaches for sentence-level sentiment analysis available today were developed only for English and there are only a few efforts that approach the problem considering other languages. Particularly, reference [6] investigates the problem of sentiment detection in three different languages: French, German, and Spanish. Their main focus is on evaluating how an automatic translation of text would work to train sentiment analysis classifiers. Similarly, Banea [7] shows a polarity classification approach, which investigates the consequence of automatic corpora generation to sentiment analysis of languages that do not have specific resources or tools. Considering automatic translation to Romanian and Spanish, they investigate the performance of polarity classification from a labeled English corpus. Although these efforts provide an important contribution, they do not cover none of the unsupervised methods evaluated in our work. They focused only on supervised machine translation learning techniques.

A recent effort [20] proposes a set of seed words (adverbs) that are expanded to train classifiers. The labeled dataset for training in several languages was automatically built considering independent language features, such as *emoticons* [31]. They conduct experiments individually and combined analysis for English, German, French and Portuguese, providing limited evaluations for specific scenarios. In the same direction, Abdel-Hady *et al.* [1] propose an unsupervised method to analyze polarity in Portuguese and Spanish, based on a language-specific resource (WordNet).

There several methods proposed to different languages including, Arabic [3], Portuguese [33], German [29], and Russian [39]. Additionally, SentiStrength [36], a known sentence-level method originally proposed and validated for English, has a version for a few other languages. In common, the above efforts focus on adapting to other language existing sentiment analysis methods and strategies to train classifiers. Overall, they provide limited baseline comparisons and validations. More important, these efforts leave a number of unanswered questions. First, to the best of our knowledge there is no effort in the literature that investigates which of the existing efforts would be more appropriate for the use of text translation strategies. Second, it is unclear if currently available specific-language strategies are able to surpass existing sentiment analysis for English if we apply text translation to English. Thus, our effort takes a different direction as it consists of providing an extensive evaluation of English sentiment analysis methods applied to translated data from different languages.

[2]http://www.internetworldstats.com/stats7.htm

3. METHODOLOGY

Our methodology to evaluate sentiment analysis in multiple languages involves three key elements. The first is a large set of sentiment analysis methods, designed for English, and commonly used for the same task (i.e. identifying if a Web 2.0 piece of text is positive or negative). To do that, we performed a large search in the literature and contacted authors to gathered a set of the "state-of-the-practice" sentiment analysis methods for English. Section 3.1 describes this effort. Second, we need to obtain a large set of labeled datasets in different languages to use as the gold standard data. We followed a similar approach of contacting several authors and, in total, we obtained datasets in nine different languages, described in Section 3.2. Finally, as a baseline for comparison we use sentiment analysis systems and tools designed for non-English text, described in Section 3.3. We call them native methods, as they were originally designed for a native non-English language.

3.1 Sentiment Analysis Methods

The term sentiment analysis has been used to describe different tasks and problems. For example, it is common to see sentiment analysis to be used to describe efforts that attempt to extract opinions from reviews [17], gauge the news polarity [28], as well as for tasks that attempt to measure mood fluctuations [16]. We restrict our focus on those efforts related to detecting the polarity (i.e., positivity or negativity) of a given text, which can be done with small adaptations on a number of existing efforts [5, 14].

Sentiment analysis also varies according to the granularity of the text, for example, document-level or sentence-level. Our focus relies on the sentence-level methods as they have been widely used in tasks related to analyzing social network data and user generated text in Web 2.0 systems.

We also note that sentence-level sentiment analysis methods can be supervised or unsupervised, which means that the former strategy requires labeled training data and the last does not. Supervised learning methods are usually very effective for sentiment analysis [9], but requires labeled data. We focus our effort on evaluating unsupervised efforts as they can be easily deployed in Web services and applications without the need of human labeling or any other intervention. Some of the methods have used machine learning to build lexicon dictionaries or even to build models and tune specific parameters. We incorporate as baseline those methods released as tools by their authors and that can be used in an unsupervised way.

Our effort to identify a high number of sentiment analysis methods

consisted of a systematically search for them in the main conferences in the field and then checking their citations and those papers that cited them. It is important to notice that some methods are available for download on the Web, others were kindly shared by their authors under request and a small part of them were reproduced from a paper that describes the method. This usually happened when authors shared only the lexical dictionaries they created, letting the implementation of the method that use the lexical resource to ourselves. Table 1 presents an overview of these methods, the reference paper in which they were published and the output format of each method. We colored as blue the outputs we consider as positive and red as negative. Black outputs are considered as neutral. As summarized in Table 1, we slightly modified some methods to adequate their output formats to the polarity detection task. We plan to release all the codes used in this paper, except for paid software like LIWC and SentiStrength, as an attempt to allow reproducibility as well as allow other efforts to question any sort of assumption and adaptation we needed to make in our experiments.

Table 1: Overview of the sentence-level methods

Methods	Output
Emoticons	-1, 1
Opinion Lexicon [17]	-1, 0, 1
Happinnes Index [12]	1, 2, 3, 4, 5, 6, 7, 8, 9
LIWC [35]	negEmo, posEmo
SenticNet [8]	negative, positive
AFFIN [24]	-1 ,0, 1
SO-CAL [34]	[<0), 0, (>0]
Emoticons DS [16]	-1, 1
NRC Hashtag [21]	sadness, anger, fear, disgust, anticipation, surprise, joy, trust
Emolex [22]	negative, positive
Umigon [19]	Negative, Neutral, Positive
Vader [18]	-1, 0, 1
PANAS-t [15]	fear, sadness, guilt, hostility, shyness, fatigue, attentiveness, joviality, assurance, serenity, surprise
Pattern.en [10]	<0.1, ≥0.1
SASA [37]	Negative, Neutral, Unsure, Positive
Stanford Rec. Deep Model [32]	very negative, negative, neutral, positive, very positive
Opinion Finder [38]	Negative, Neutral, Positive
SentiWordNet [13]	-1, 0, 1
SANN [26]	neg, neu, pos
Sentiment140 [23]	Negative, Neutral, Positive
SentiStrength [36]	-1, 0, 1

3.2 Gold Standard Datasets

In this section, we present an overview of the datasets used in this work. These datasets consist of eleven gold standard datasets of short messages, which were labeled by humans as positive or negative according to their sentiment polarity.

These datasets consist of data in 9 different languages, besides two sets of messages in English. Their content are from different contexts from Twitter and Website reviews. Smaller datasets contain dozens of instances and some of them few thousands of posts. Random tweets include data of different subjects and the review datasets consist of labeled messages from costumers reviews about different products and movies. To allow a fair comparison, we selected messages in English from these two groups of data, *Random tweets* and *Reviews*. Table 2 summarizes the relevant information about these datasets[3].

[3]The datasets used in this paper are public available at http://www.dcc.ufmg.br/~fabricio.

Table 2: Gold standard labeled datasets

Dataset Language	Description	# Pos.	# Neg.
Arabic	Tweets in Arabic [3]	1,000	1,000
Dutch	Tweets in Dutch[4]	88	63
French	Tweets in French [31]	159	160
German	Tweets in German [31]	143	95
Italian	Tweets in Italian[5]	820	1,422
Portuguese	Tweets in Portuguese [31]	297	213
Russian	Tweets in Russian[6]	1,145	1,188
Spanish	Tweets in Spanish[7]	683	350
English Twitter	Tweets in English [18]	2,897	1,299
Turkish	Reviews in Turkish [11]	5,600	2,800
English Reviews	Reviews in English [18]	2,128	1,482

3.3 Language-Specific Methods

Ideally, we would like to compare the use of machine translation using all the methods designed for English described in Section 3.1 with a large number of methods proposed for some specific language. We contacted authors of some identified efforts asking for datasets and their methods. While we succeeded in obtaining a large number of datasets, most of these methods are not available even under request to authors, making reproducibility almost impossible in most of the cases. We were able to assess a Multilanguage version of Sentistrength (ML-Sentistrength), available in Dutch, French, German, Italian, Portuguese, Spanish, and Turkish. Our second baseline is a commercial sentiment analysis API namely Semantria (SMTR)[8], which provides results in French, German, Italian, Portuguese and Spanish. We used the trial version of the Microsoft Excel Plugin available on their website.

4. EXPERIMENTAL RESULTS

This section presents our experimental results. First we describe the evaluation metrics used along the rest of the paper.

4.1 Evaluation Metrics

The metrics used to evaluate the methods were based on *precision* and *recall* to obtain the *F1* per class (positive and negative). $F1$ measure is the harmonic mean between both precision and recall. As some of our datasets are unbalanced we use Macro-F1 instead of accuracy to summarize the overall prediction performance of each method in each dataset. Macro-F1 values are computed by first calculating F1 values for each class in isolation, as exemplified above for negative, and then averaging over all classes. Thus, Macro-F1 considers equally important the effectiveness in *each class*, independently of the relative size of the class.

Coverage is the fraction of messages that a method is able to classify as either positive or negative in a given dataset. Ideally, polarity detection methods should retain high coverage to avoid bias in the results, due to the unidentified messages. For instance, suppose that a sentiment method has classified only 10% of a given set of tweets. The remaining 90% consisting of unidentified tweets may completely change the result, that is, whether the context drawn from tweets should be positive or negative. Therefore, having high coverage in data is essential in analyzing Web data.

As we have a large number of methods and datasets to compare, we consider a metric proposed in [27], called *Winning Number*. This measure assess the most competitive methods among a series of candidates, given a large series of predefined tasks they

[8]https://semantria.com

have to perform. That is, the *Winning Number* of a method i in the context of a performance measure M, is given as $S_i(M) = \sum_{j=1}^{5} \sum_{k=1}^{23} \mathbf{1}_{M_i(j)>M_k(j)}$, where j is the dataset index (5 datasets that all methods can be evaluated) , i and k are the methods' index (23 methods), $M_i(j)$ is the performance of the $i-th$ method on $j-th$ dataset in terms of measure M, and $\mathbf{1}_{M_i(j)>M_k(j)}$ is the indicator function:

$$\mathbf{1}_{M_i(j)>M_k(j)} = \begin{cases} 1 & \text{if } M_i(j) > M_k(j), \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the larger $S_i(M)$ is, the better the $i-th$ method performs compared to the others. The results of *Winning Number* of Coverage and Macro-F1 are shown in Table 4.

4.2 English vs. Non-English

We begin by comparing the performance of all methods for English and all non-English datasets. Figure 1 compares the Macro-F1 results of the execution all methods between each language to an English reference. The bottom of the bar corresponds to the minimum value, then the first quartile, the square inside the bar represents the median, then the third quartile and the top of the bar corresponds to the maximum value.

As can be observed from these results, the overall Macro-F1 for random tweets (blue) datasets in different languages mostly appear between 0.6 and 0.8 including English. All highest performances are above 0.8 and lowest performances on each language are close to 0.4. For Dutch, we have a very low minimum result due to lack of emoticons. Moreover, we have the results for product reviews (yellow). Although they differ from tweets performance, Turkish follows the English performance, where the lowest Macro-F1 are close to 0.3 and the first quartile are close to 0.5, although English has a better performance above the first quartile. An exception is Arabic, where the difference in the median result compared to English is 0.375. We believe this is related to the morphological richness of Arabic [2], i.e., a significant amount of information concerning syntactic units and relations is expressed at the word-level and this could be difficult to machine translation. We also note the high difference on the performance of the methods on datasets of tweets and product reviews, suggesting that the methods we consider in this paper are mostly appropriated to the informal text of tweets.

Overall, these results suggest that machine translation leads to a lower prediction performance in comparison with English, but it also suggests that machine translation can be a competitive strategy if the suitable sentiment analysis method is chosen for the task. Next, we detail the performance of each method across all datasets.

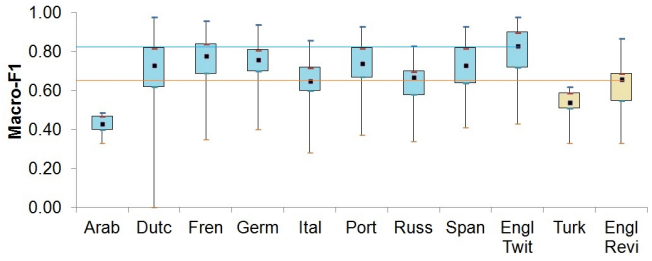

Figure 1: Macro-F1 across languages

4.3 Overall Prediction Performance

Next, we compare the results among the 21 English sentiment analysis methods in 9 translated non-English datasets. We also compare results with two language-specific commercial approaches presented in Section 3.3.

Table 4.3 presents the performance of F1 per class as well as Macro-F1 and Coverage of these methods. As our first observation, we identify a strong variation on the prediction performances of some methods for each different languages. For example, the LIWC obtained a Macro-F1 of 0.88 for the translated French dataset, which is much better than the 0.52 obtained for the Spanish dataset. We can also note that emoticons showed to be a good method for detecting positive and negative messages when the input data has an emoticon. However, it considers most of the instances as neutral, leading to a bad performance in terms of coverage for most of the datasets. Thus, one needs to consider the tradeoff between coverage and Macro-F1 to select the appropriate method for multilanguage sentiment analysis based on machine translation. We note that some methods obtain consistent results for Macro-F1 still keeping high values of coverage across multiple languages, such as SentiStrengh, Umigon, SO-CAL, and Vader, although Vader presented quite low coverage values for Turkish and the English datasets. This suggests that these methods might be the most reliable ones for machine translation in the languages analyzed.

Finally, we noted by the analysis of the $F1$ scores that most methods are more accurate in correctly classifying positive than negative text, suggesting that methods can lead to bias in their analysis towards positivity.

4.4 Language-Specific Methods

When we look at the two commercial language-specific tools, ML sentistrength and Semantria, we can note that ML sentistrength has the *2th* best Macro-F1 for Germany behind only emoticons (which has the worst coverage), but it showed a very low Macro-F1 for Turkish data (0.47), appearing only in the *19th* position in the ranking for that dataset. In Dutch, ML sentistrength has a poor performance (0.58) in comparison to its original English Sentistrength implementation, which is the best method with 0.97 for the English dataset of tweets. This same situation repeats along the results, suggesting that the original Sentistrength is better with machine-translated text than the language-specific sentistrength. Overall, this suggests that machine translation across the methods and datasets we evaluate in our effort could become a baseline for comparison of any novel language specific method. Overall, Semantria was able to analyze 51% of the data and has 0.82 of Macro-F1, that is, in average, 0.09 below the best method of each dataset. Although Semantria is not the best method in any dataset, it has a persistent position among the top methods.

4.5 Ranking Methods

Finally, Table 4 presents the Winning Points, considering results of French, Germany, Italian, Portuguese and Spanish datasets, which are the languages in which all methods could be evaluated. From this analysis, we obtain a general rank of the performance of the methods. As we can see, Sentistrength has the highest rank position for Macro-F1, even better than language-specific methods. In the other hand, it is the *20th* one in terms of coverage. The language-specific methods do not present the best results, although Semantria has the *3th* and Multi-Language Sentistrength the *6th* position, very close to Umigon(*4th*) and SO-CAL(*5th*). Finally, we

Table 3: Prediction performance for all methods across all datasets.

	Arabic				Dutch				French				Germany			
Method	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov
AFINN	0.65	0.11	0.38	0.56	0.86	0.78	0.82	0.58	0.85	0.77	0.81	0.54	0.87	0.70	0.79	0.50
Emolex	0.61	0.31	0.46	0.53	0.79	0.71	0.75	0.63	0.70	0.75	0.73	0.55	0.76	0.67	0.71	0.49
Emoticons	0.67	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.96	0.80	0.88	0.09	0.97	0.91	0.94	0.20
Emoticons DS	0.67	0.00	0.33	0.99	0.74	0.00	0.37	1.00	0.67	0.03	0.35	0.98	0.75	0.04	0.40	0.99
Happinnes Index	0.66	0.07	0.36	0.67	0.78	0.48	0.63	0.72	0.76	0.56	0.66	0.58	0.80	0.45	0.63	0.58
LIWC	0.64	0.15	0.40	0.58	0.90	0.78	0.84	0.68	0.90	0.85	0.88	0.63	0.89	0.73	0.81	0.59
NRC Hashtag	0.57	0.40	0.49	0.82	0.49	0.72	0.61	0.68	0.56	0.76	0.66	0.73	0.65	0.74	0.70	0.72
Opinion Finder	0.60	0.33	0.47	0.36	0.87	0.87	0.87	0.46	0.77	0.76	0.76	0.45	0.79	0.72	0.76	0.41
Opinion Lexicon	0.63	0.22	0.43	0.48	0.85	0.78	0.81	0.58	0.80	0.78	0.79	0.56	0.90	0.78	0.84	0.49
PANAS-t	0.55	0.44	0.49	0.03	0.71	0.76	0.73	0.13	0.96	0.97	0.96	0.08	0.78	0.67	0.72	0.08
Pattern.en	0.64	0.19	0.42	0.52	0.89	0.81	0.85	0.72	0.84	0.77	0.81	0.72	0.88	0.73	0.81	0.79
SANN	0.63	0.24	0.43	0.50	0.77	0.69	0.73	0.50	0.81	0.74	0.78	0.53	0.82	0.72	0.77	0.46
SASA	0.58	0.39	0.48	0.61	0.74	0.48	0.61	0.57	0.71	0.51	0.61	0.60	0.74	0.55	0.64	0.61
SenticNet	0.65	0.10	0.38	0.90	0.77	0.49	0.63	0.91	0.76	0.61	0.69	0.92	0.81	0.58	0.70	0.89
Sentiment140	0.60	0.33	0.47	0.91	0.73	0.63	0.68	0.99	0.74	0.73	0.73	0.96	0.76	0.72	0.74	0.98
SentiStrength	0.63	0.22	0.43	0.26	0.98	0.98	0.98	0.32	0.95	0.95	0.95	0.34	0.93	0.86	0.89	0.27
SentiWordNet	0.63	0.23	0.43	0.84	0.72	0.52	0.62	0.91	0.76	0.65	0.70	0.86	0.73	0.57	0.65	0.89
SO-CAL	0.62	0.26	0.44	0.61	0.83	0.75	0.79	0.77	0.84	0.83	0.84	0.69	0.86	0.76	0.81	0.65
Stanford Deep Mode	0.58	0.39	0.48	0.66	0.51	0.70	0.61	0.91	0.56	0.74	0.65	0.86	0.69	0.68	0.68	0.84
Umigon	0.62	0.29	0.45	0.43	0.82	0.78	0.80	0.70	0.87	0.85	0.86	0.64	0.90	0.82	0.86	0.69
Vader	0.65	0.15	0.40	0.81	0.86	0.77	0.82	0.85	0.86	0.81	0.83	0.82	0.86	0.72	0.79	0.80
ML SentiStrength	-	-	-	-	0.67	0.50	0.58	0.03	0.86	0.76	0.81	0.09	0.90	0.93	0.92	0.15
Semantria	-	-	-	-	-	-	-	-	0.85	0.85	0.85	0.60	0.92	0.83	0.88	0.43

	Italian				Portuguese				Russian				Spanish			
Method	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov
AFINN	0.70	0.72	0.71	0.56	0.86	0.79	0.82	0.51	0.70	0.67	0.68	0.47	0.90	0.78	0.84	0.58
Emolex	0.54	0.78	0.66	0.64	0.66	0.69	0.68	0.53	0.54	0.69	0.61	0.54	0.72	0.63	0.68	0.59
Emoticons	0.88	0.67	0.77	0.04	0.93	0.86	0.90	0.06	0.91	0.67	0.79	0.07	0.94	0.50	0.72	0.07
Emoticons DS	0.54	0.01	0.28	0.99	0.74	0.00	0.37	0.98	0.66	0.02	0.34	0.97	0.80	0.02	0.41	0.99
Happinnes Index	0.62	0.50	0.56	0.50	0.80	0.54	0.67	0.57	0.67	0.43	0.55	0.56	0.84	0.48	0.66	0.57
LIWC	0.60	0.59	0.60	0.61	0.79	0.65	0.72	0.61	0.66	0.62	0.64	0.55	0.73	0.30	0.52	0.63
NRC Hashtag	0.46	0.78	0.62	0.82	0.50	0.66	0.58	0.65	0.39	0.71	0.55	0.82	0.55	0.60	0.57	0.69
Opinion Finder	0.68	0.77	0.73	0.40	0.77	0.78	0.77	0.42	0.63	0.71	0.67	0.40	0.84	0.75	0.80	0.39
Opinion Lexicon	0.69	0.75	0.72	0.50	0.84	0.79	0.81	0.44	0.69	0.71	0.70	0.44	0.88	0.77	0.82	0.52
PANAS-t	0.67	0.75	0.71	0.05	0.77	0.72	0.74	0.08	0.64	0.73	0.69	0.08	0.91	0.76	0.84	0.05
Pattern.en	0.68	0.62	0.65	0.64	0.86	0.74	0.80	0.72	0.74	0.64	0.69	0.65	0.89	0.62	0.76	0.68
SANN	0.65	0.65	0.65	0.46	0.82	0.75	0.78	0.46	0.70	0.64	0.67	0.50	0.89	0.72	0.81	0.44
SASA	0.56	0.54	0.55	0.62	0.77	0.61	0.69	0.58	0.56	0.63	0.60	0.61	0.78	0.36	0.57	0.52
SenticNet	0.58	0.45	0.51	0.94	0.78	0.54	0.66	0.91	0.66	0.46	0.56	0.88	0.82	0.46	0.64	0.93
Sentiment140	0.57	0.67	0.62	0.97	0.75	0.68	0.71	0.96	0.58	0.64	0.61	0.98	0.81	0.65	0.73	0.98
SentiStrength	0.84	0.88	0.86	0.26	0.94	0.91	0.93	0.31	0.80	0.85	0.83	0.24	0.95	0.90	0.93	0.29
SentiWordNet	0.59	0.59	0.59	0.89	0.76	0.59	0.67	0.89	0.63	0.53	0.58	0.88	0.80	0.57	0.69	0.91
SO-CAL	0.70	0.78	0.74	0.67	0.85	0.81	0.83	0.66	0.70	0.74	0.72	0.60	0.89	0.80	0.85	0.70
Stanford Deep Mode	0.46	0.81	0.64	0.89	0.51	0.67	0.59	0.86	0.45	0.72	0.58	0.85	0.46	0.59	0.53	0.91
Umigon	0.76	0.79	0.78	0.53	0.87	0.82	0.84	0.56	0.75	0.77	0.76	0.54	0.90	0.71	0.80	0.52
Vader	0.70	0.73	0.72	0.80	0.87	0.80	0.83	0.79	0.73	0.71	0.72	0.75	0.90	0.79	0.84	0.82
ML SentiStrength	0.63	0.67	0.65	0.02	0.95	0.83	0.89	0.05	-	-	-	-	0.86	0.79	0.82	0.07
Semantria	0.67	0.69	0.68	0.46	0.87	0.82	0.85	0.56	-	-	-	-	0.90	0.84	0.87	0.54

	Turkish				English Twitter				English Reviews			
Method	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov	F1(+)	F1(-)	Macro-F1	Cov
AFINN	0.69	0.32	0.51	0.82	0.93	0.93	0.93	0.64	0.82	0.57	0.69	0.49
Emolex	0.59	0.59	0.59	0.89	0.75	0.82	0.78	0.57	0.54	0.55	0.55	0.61
Emoticons	0.84	0.20	0.52	0.01	0.99	0.98	0.98	0.12	0.00	0.67	0.33	0.00
Emoticons DS	0.67	0.00	0.33	1.00	0.70	0.16	0.43	0.95	0.75	0.02	0.38	0.97
Happinnes Index	0.66	0.13	0.40	0.93	0.77	0.68	0.72	0.60	0.78	0.25	0.52	0.47
LIWC	0.69	0.33	0.51	0.83	0.82	0.53	0.68	1.00	0.68	0.30	0.49	0.97
NRC Hashtag	0.39	0.69	0.54	0.52	0.62	0.79	0.70	0.70	0.48	0.65	0.56	0.69
Opinion Finder	0.58	0.57	0.58	0.83	0.81	0.88	0.84	0.44	0.73	0.60	0.67	0.37
Opinion Lexicon	0.70	0.40	0.55	0.83	0.88	0.89	0.89	0.55	0.82	0.61	0.72	0.55
PANAS-t	0.68	0.50	0.59	0.12	0.91	0.91	0.91	0.10	0.91	0.47	0.69	0.03
Pattern.en	0.70	0.34	0.52	0.97	0.88	0.87	0.88	0.68	0.81	0.55	0.68	0.70
SANN	0.68	0.44	0.56	0.85	0.83	0.84	0.83	0.51	0.80	0.52	0.66	0.44
SASA	0.63	0.43	0.53	0.74	0.70	0.68	0.69	0.59	0.70	0.49	0.59	0.66
SenticNet	0.66	0.11	0.39	0.99	0.80	0.72	0.76	0.87	0.75	0.32	0.54	0.91
Sentiment140	0.66	0.50	0.58	1.00	0.76	0.76	0.76	0.95	0.61	0.60	0.61	0.99
SentiStrength	0.78	0.40	0.59	0.39	0.97	0.98	0.97	0.35	0.94	0.79	0.87	0.20
SentiWordNet	0.66	0.36	0.51	1.00	0.60	0.67	0.64	0.92	0.69	0.47	0.58	0.92
SO-CAL	0.70	0.53	0.62	0.95	0.88	0.90	0.89	0.67	0.83	0.70	0.76	0.72
Stanford Deep Mode	0.51	0.68	0.59	0.94	0.65	0.79	0.72	0.87	0.68	0.69	0.69	0.80
Umigon	0.58	0.61	0.60	0.87	0.89	0.91	0.90	0.67	0.76	0.67	0.71	0.51
Vader	0.80	0.22	0.51	0.01	0.98	0.97	0.98	0.26	0.91	0.76	0.84	0.07
ML SentiStrength	0.70	0.24	0.47	0.02	-	-	-	-	-	-	-	-
Semantria	-	-	-	-	-	-	-	-	-	-	-	-

highlight the trade-off between Coverage and Macro-F1 exposed by Emoticons DS, which is the worst method in Macro-F1 but the best in Coverage, such trade-off can also be verified for Emoticons and Sentistrength methods. SO-CAL, Umigon, and Vader are methods that are able to perform well in terms of Macro-F1 and Coverage in most of the datasets.

5. CONCLUSION

Sentiment analysis is emerging as a key tool for social media analysis. As most of the existing sentiment analysis methods were designed to English, it is crucial to develop new technologies able to leverage sentiment analysis to a wide number of other languages. In this work, we provide an extensive evaluation of machine translation for sentiment analysis, considering 21 English sentence-level methods across datasets in 9 different languages. We also compare those evaluations with 2 popular commercial language-specific approaches. As a result, Sentistrength showed to be the most accurate method for the task of detecting sentiments using machine translation. However, we also identify SO-CAL, Umigon, and Vader as methods that are able to perform well in different datasets, considering Macro-F1 and Coverage.

Our findings suggest that machine translation leads to a lower prediction performance for non-English data in comparison with text in English, but it also suggests that machine translation can be a competitive strategy if the suitable sentiment analysis method is properly chosen. Additionally, we show that two popular language-specific methods do not have a significant advantage over a ma-

Table 4: Winning Numbers

Method	Macro-F1	Method	Coverage
SentiStrength	107	Emoticons DS	110
Emoticons	92	Sentiment140	105
Semantria	89	SenticNet	100
Umigon	87	SentiWordNet	94
SO-CAL	87	Stanford Deep Mode	91
ML. SentiStrength	82	Vader	84
Vader	76	NRC Hashtag	77
PANAS-t	74	Pattern.en	74
Opinion Lexicon	74	SO-CAL	73
AFINN	73	LIWC	60
Opinion Finder	61	Umigon	52
Pattern.en	59	SASA	52
SANN	57	Emolex	48
LIWC	50	Happinnes Index	47
Sentiment140	40	AFINN	43
Emolex	40	**Semantria**	41
SentiWordNet	25	Opinion Lexicon	37
NRC Hashtag	20	SANN	27
SenticNet	19	Opinion Finder	20
Happinnes Index	19	SentiStrength	15
Stanford Deep Mode	18	Emoticons	7
SASA	16	PANAS-t	4
Emoticons DS	0	**ML. SentiStrength**	4

chine translation approach. For example, even the original sentistrength with machine-translated text showed to be better than the language-specific one. Given the simplicity that the strategy that machine translation offers, one may prefer to deploy it at some cost on the prediction performance instead of developing a solution on each specific language.

As a final contribution, we note that obtaining labeled datasets in different languages and implement a high number of sentiment analysis methods is a very labor task. We plan to release to the scientific community all the methods codes and labeled datasets used in this paper hoping that it can help sentiment analysis to become English independent. We hope that machine translation across the methods and datasets we evaluate in our effort could become a baseline for comparison of any novel language specific method.

Acknowledgements

This research is funded by grants from CNPq, CAPES and FAPEMIG.

6. REFERENCES

[1] M. Abdel-Hady, R. Mansour, and A. Ashour. Cross-lingual twitter polarity detection via projection across word-aligned corpora. In *Proc. of WISDOM*, 2014.

[2] M. Abdul-Mageed, M. T. Diab, and M. Korayem. Subjectivity and sentiment analysis of modern standard arabic. In *Proc. of ACL-HLT*, 2011.

[3] N. Abdulla, N. Ahmed, M. Shehab, and M. Al-Ayyoub. Arabic sentiment analysis: Lexicon-based and corpus-based. In *Proc. of AEECT, IEEE*, 2013.

[4] M. Aiken and S. Balan. An analysis of google translate accuracy. *Translation journal*, 16(2):1–3, 2011.

[5] M. Araújo, P. Gonçalves, F. Benevenuto, and M. Cha. ifeel: A system that compares and combines sentiment analysis methods. In *Proc. of WWW*, 2014.

[6] A. Balahur and M. Turchi. Multilingual sentiment analysis using machine translation? In *Proc. of WASSA*, 2012.

[7] C. Banea, R. Mihalcea, J. Wiebe, and S. Hassan. Multilingual subjectivity analysis using machine translation. In *Proc. of EMNLP*, 2008.

[8] E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *Proc. of AAAI Fall Symposium Series*, 2010.

[9] S. Canuto, M. A. Gonçalves, and F. Benevenuto. Exploiting new sentiment-based meta-level features for effective sentiment analysis. In *Proc. of WSDM*, 2016.

[10] T. De Smedt and W. Daelemans. Pattern for python. *The Journal of Machine Learning Research*, 13(1):2063–2067, 2012.

[11] E. Demirtas and M. Pechenizkiy. Cross-lingual polarity detection with machine translation. In *Proc. of WISDOM*, 2013.

[12] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2009.

[13] Esuli and Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proc. of LREC*, 2006.

[14] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *Proc. of COSN*. ACM, 2013.

[15] P. Gonçalves, F. Benevenuto, and M. Cha. PANAS-t: A Pychometric Scale for Measuring Sentiments on Twitter. abs/1308.1857v1, 2013.

[16] A. Hannak, E. Anderson, L. F. Barrett, S. Lehmann, A. Mislove, and M. Riedewald. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Proc. of ICWSM*, 2012.

[17] M. Hu and B. Liu. Mining and summarizing customer reviews. In Proc. of KDD, 2004.

[18] C. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. of ICWSM*, 2014.

[19] C. Levallois. Umigon: sentiment analysis for tweets based on lexicons and heuristics. In *Proc. of SemEval*, 2013.

[20] Z. Lin, S. Tan, and X. Cheng. Language-independent sentiment classification using three common words. In *Proc. of CIKM*, 2011.

[21] S. Mohammad. #emotional tweets. In *Proc. of SemEval*, 2012.

[22] S. Mohammad and P. D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[23] S. M. Mohammad, S. Kiritchenko, and X. Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proc. of SemEval*, 2013.

[24] F. Å. Nielsen. A new anew: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*, 2011.

[25] N. Online. Social networks and blogs now 4th most popular online activity, ahead of personal email, nielsen reports. http://www.nielsen.com/us/en/press-room/2009/social_networks__.html, 2009. Accessed in April, 08, 2015.

[26] N. Pappas and A. Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proc. of SIGIR*, 2013.

[27] T. Qin, T.-Y. Liu, J. Xu, and H. Li. Letor: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4), Aug. 2010.

[28] J. Reis, F. Benevenuto, P. Vaz de Melo, R. Prates, H. Kwak, and J. An. Breaking the news: First impressions matter on online news. In *Proc. of ICWSM*, 2015.

[29] R. Remus, U. Quasthoff, and G. Heyer. Sentiws-a publicly available german-language resource for sentiment analysis. In *Proc. of LREC*, 2010.

[30] F. Ribeiro, M. Araújo, P. Gonçalves, F. Benevenuto, and M. A. Gonçalves. A benchmark comparison of state-of-the-practice sentiment analysis methods. *arXiv preprint arXiv:1512.01818*, 2015.

[31] M. H. Sascha Narr and S. Albayrak. Language-independent twitter sentiment analysis. In *Proc. of KDML*, 2012.

[32] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*, 2013.

[33] M. Souza and R. Vieira. Sentiment analysis on twitter data for portuguese language. In *Proc. of PROPOR*. 2012.

[34] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, 2011.

[35] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *J. of Lang. and Soc. Psych.*, 2010.

[36] M. Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf, 2013. Accessed in August, 22, 2015.

[37] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proc. of ACL*, 2012.

[38] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proc. of HLT/EMNLP*, 2005.

[39] N. Yussupova, D. Bogdanova, and M. Boyko. Applying of sentiment analysis for texts in russian based on machine learning approach. In *Proc. of IMMM*, 2012.