

Multi-Lingual Sentiment Analysis of Social Data Based on Emotion-Bearing Patterns

Carlos Argueta

National Tsing Hua University
No. 101, Section 2, Kuang-Fu Road
Hsinchu, Taiwan
kid.a.rgueta@gmail.com

Yi-Shin Chen

National Tsing Hua University
No. 101, Section 2, Kuang-Fu Road
Hsinchu, Taiwan
yishin@gmail.com

Abstract

Social networking sites have flooded the Internet with posts containing shared opinions, moods, and feelings. This has given rise to a new wave of research to develop algorithms for emotion detection and extraction on social data. As the desire to understand how people feel about certain events/objects across countries or regions grows, the need to analyze social data in different languages grows with it. However, the explosive nature of data generated around the world brings a challenge for sentiment-based information retrieval and analysis. In this paper, we propose a multilingual system with a computationally inexpensive approach to sentiment analysis of social data. The experiments demonstrate that our approach performs an effective multi-lingual sentiment analysis of microblog data with little more than a 100 emotion-bearing patterns.

1 Introduction

Web 2.0 and the rise of social networking platforms such as microblogs have brought new ways to share opinions in a global setting. Microblogging sites represent a new way to share information about everything, such as new products, places of interest or popular culture in some ways replacing traditional word-of-mouth communication. Those sites have become rich repositories of opinions from audiences diverse in culture, race, location, and language. They represent, for people and businesses, a potential opportunity to understand what the global community thinks about them, helping them to make better informed decisions when improving their image and products. They may also offer the general public a way to find useful information and opinions before purchasing a product or service.

Vast swathes of the global population have access to nearly the same products and services. For that reason, being aware of opinions from around the world, regardless of the languages, is no longer ambitious but necessary. Systems which are able to aggregate opinionated data in multiple languages could highlight a global trend around a target query. This is desirable as targets may have different impacts depending on the area. The possibilities are huge but the challenges are many. As in every major endeavor, it is necessary to start with the basics, such as language detection and sentiment analysis.

With the explosive nature of subjective data in the Web, several patterns analysis techniques such as the ones by Yi et al. (2003) and Davidov et al. (2010) have been proposed to extract opinionated and emotion-bearing data. Most works have focused on English language, except the technique proposed in (Sascha Narr and Albayrak, 2012) which utilizes n-grams as language independent features to classify tweets by their polarity in 4 different languages. However, since this technique relies solely on frequency statistics, it would need large training datasets.

In this paper, we propose a language independent approach to emotion-bearing patterns retrieval in microblog data. Each extracted pattern and its related words can also be considered as n-grams, however, selected in a way that makes them more semantically related to the domain of sentiment and emotion analysis. The proposed multi-lingual framework consists of two stages: Filter and Refine approach. In the Filter stage, the language and a hint of the polarity of a microblog post are detected based on the

This work is licenced under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: <http://creativecommons.org/licenses/by/4.0/>

n-gram classification approach. Our approach is different from the one proposed in (Sascha Narr and Albayrak, 2012) in that the n-grams are used at the character-level instead of the word-level.

In the Refine stage, language-independent emotion-bearing words and patterns (using an adaptation of extraction methods by Davidov and Rappoport (2006)) are employed. Our approach differs from the method of Davidov et al. (2010) in that we use words belonging to empirically defined psychological and structural categories (emotion, cognition, etc.)¹. These words are used as seeds to restrict the features extracted to emotion bearing patterns and words.

Our experimental results show that our approach can extract relevant patterns. With only about 100 patterns employed of about 31,500 n-gram candidates, the approach presented in this paper outperforms a state-of-the-art classifier for the French language. These results validate the potential of the proposed technique.

2 Related Work

Patterns have been extensively used for many Information Retrieval tasks. Pantel and Pennacchiotti (2006) extract semantic relations using generic patterns (with broad coverage but low precision). The key assumption is that in very large corpora like the Web, correct instances generated by a generic pattern will be instantiated by some reliable patterns, where reliable patterns are patterns that have high precision but often very low recall. Wang et al. (2007) propose a topical n-gram (TNG) model that automatically determines unigram words and phrases based on context, and assigns a mixture of topics to both individual words and n-gram phrases. They present an Information Retrieval (IR) application with better performance on an ad-hoc retrieval task over a TREC collection. Davidov and Rappoport (2006) have proposed an approach to unsupervised discovery of word categories based on symmetric patterns. A symmetric pattern is one in which co-occurring words can potentially belong to the same semantic category.

With the speed at which subjective data is generated on the Web, and its potential usage, patterns have also been applied to extract opinionated data. Yi et al. (2003) use a Sentiment Pattern Database to match sentiment phrases related to a subject from online texts. The system was verified using online product review articles. Dave et al. (2003) tried statistical and linguistic substitutions to transform specific patterns into more general ones. They also used an algorithm based on suffix trees to determine substrings that provide optimal classification. The extracted features were applied to opinion extraction of product reviews. In their work, Davidov et al. (2010) use their pattern definition (Davidov and Rappoport, 2006) to identify diverse sentiment types in short microblog texts.

Most of the work with patterns for sentiment extraction has focused on English language. Although, a few studies have identified patterns as a tool for bridging the gap between languages. Cui et al. (2011) automatically extracted different types of word level patterns (denoted emotion tokens), and labeled their sentiment polarities with an unsupervised propagation algorithm. Sascha Narr and Albayrak (2012) utilized n-grams as language independent features to classify tweets by their polarity in 4 different languages.

The main drawback of features such as n-grams is that, to capture the semantics of a specific domain, they rely solely on frequency statistics. This impacts less when large training data is available, but can be a considerable disadvantage when the data is scarce. Another drawback of such features is the large number of n-grams that need to be included in the features set. Larger feature spaces have a big impact on the efficiency of systems.

3 Methodology

The objective of this work is to propose an effective multi-lingual emotion-bearing patterns extraction approach. To test the relevance of the extracted features, a multi-lingual sentiment analysis system for microblog data is defined. The proposed framework illustrated in Figure 1 consists of two stages: the *Filter stage* and the *Refine stage*. Given a set of microblog posts containing a query term, the Filter stage utilizing n-gram patterns first detects the language of all the microblog posts. Then it obtains the polarity

¹See the study by James W. Pennebaker and Booth (2007) and Tausczik and Pennebaker (2010)

(negative, positive) by utilizing a classifier trained with n-gram features at the character level. The Refine stage utilizing symmetric patterns performs a finer analysis of the posts to classify the ones that the Filter stage left out. It utilizes extracted emotion-bearing patterns as described in Section 3.1 and related words as features.

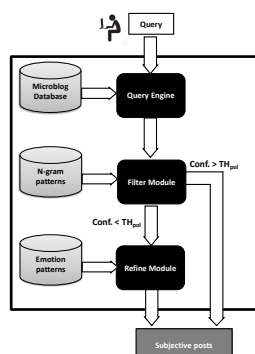


Figure 1: System Overview

3.1 Candidate Emotion-Bearing Surface Patterns Construction

The intuition of the construction method is based on psychological studies (James W. Pennebaker and Booth, 2007) proposing that emotional cues can be found in a person’s utterances and texts, providing windows into their emotional and cognitive worlds. Based on this idea, the following method that can capture the textual emotion cues in the form of patterns and words is introduced.

The proposed extraction technique in this paper adapted the unsupervised word categories discovery technique introduced by Davidov and Rappoport (2006). To guarantee that the retrieved features are relevant to the sentiment analysis task here presented, the proposed emotion-bearing pattern extraction technique requires the following adaptations.

Other than the HW and CW word types introduced by Davidov and Rappoport (2006), the psychological word type (PW) is introduced as seeds for extracting more emotion-related patterns. Words in PW are content-bearing, and pertain to psychological categories related to emotion, cognition, affection, social, perception, etc. The PW set should include words like, “peace”, “abandon”, and “dream”. By the symmetric nature of the patterns, the PW words can be naturally expanded from the extraction dataset. In order to drive the extraction towards more emotion-related patterns, it is also enforced that at least one of the CW words of a pattern contains a word from the PW set. For example, “peace is a dream” is an instance of a pattern “CW HW HW CW”.

The result of the extraction is a large list of word subsequences related to different emotion-bearing patterns. There are two main reasons not to employ all extracted subsequences from a large corpus as features. First, the current list is usually huge (around 150,000 unique subsequences found in the experiments with English data), making the training process inefficient. Second, it is necessary to account for the fact that the dataset used for features extraction may be completely different from the training, and testing sets. Additionally, the PW set can also vary greatly in size and psychological categories included. This can potentially impact the coverage/accuracy due to unseen words.

3.2 Graph-Based Relevant Patterns Selection

The following reduction method based on the graph representation for patterns described in (Davidov and Rappoport, 2006) is introduced to reduce the features space, and to account for unseen words. First, infrequent subsequences are ignored. Then, the subsequences are grouped by pattern based on their HW words. Subsequences having the same HW word in the same positions are grouped together and their CW words are replaced by “*”.² Finally, as proposed by Davidov and Rappoport (2006), the directed

²“*” is a wild card that matches any word

graph $G(p) = (V(G), E(G))$ for each pattern p is constructed. For example, subsequences “love my niece” and “hate my boss” from the meta-pattern “CW HW CW” are both grouped together as part of the pattern “* my *”.

Two different scores are proposed to measure the degree of emotion expressed by each pattern. The intuition of the following definitions is, the higher the proportion of PW words appearing in subsequences belonging to a given pattern, the higher the degree of emotion of this pattern.

Definition 1 (SC1)

The number of out-links of the vertex with max. number of out-links in the graph for pattern p .

$$SC1 = \max(|\{(x, y) | x \in V(G(p))\}|) \tag{1}$$

Definition 2 (SC2)

The number of in-links of the vertex with max. number of in-links in the graph for pattern p .

$$SC2 = \max(|\{(x, y) | y \in V(G(p))\}|) \tag{2}$$

Patterns with high SC1 and SC2 scores ensure a good coverage as they capture more PW words from the corpus. For each score, a ranked list of patterns is obtained. Patterns in top TH_{top} in at least one list and not in bottom TH_{bottom} in any list are retained. The final list of features is composed of all the retained patterns, and all the CWs it captured from the corpus through its related subsequences. This approach significantly reduces the features space by using only the relevant patterns, and not the subsequences (n-grams) as features. It also helps the coverage, as unseen words can still be captured by a pattern through the “* ” wild card.

3.3 Multi-lingual Sentiment Analysis on Microblog Data

The following Filter and Refine approach is employed to determine the polarity of posts from microblog data. Given a post, the Filter stage firsts detects its language. It does so by training language models using n-grams at the character-level. Next, the corresponding polarity models are used to determine the polarity of the post. The polarity models are also constructed with the same method.

A confidence value is obtained as follows.

$$Conf(p) = \left| \sum (Freq_{positive}(ng) - Freq_{negative}(ng)) \right| \tag{3}$$

where ng is an n-gram entry and $Freq$ gets its frequency count in a polarity model.

We define experimentally two thresholds TH_{pol} for $pol \in \{positive, negative\}$. For a given detected class pol we say the post has that polarity only if $Conf(p) > TH_{pol}$ else we send the post to the Refine stage.

The posts not classified by the Filter stage are processed by the Refine stage, which is based on a Multinomial Naïve Bayes classifier. The classifier uses the emotion-patterns and words extracted using the methods described in 3.1 and 3.2 as binary features. Each pattern is used as a regular expression to look for matching subsequences in a post. If a match exists, the corresponding value in the vector is set to 1 and to 0 otherwise. The same process applies for word features.

4 Experiments

4.1 Experimental Setup

Two experiments were performed to validate the approach introduced in this paper. Datasets in three target languages (English, Spanish, and French) were utilized to test the approach. To avoid favoring the proposed method by having similar data characteristics in both training and testing stages, different datasets are employed in the experiments.

The larger set, Set_{HT} , was collected using emotion-bearing hashtags as noisy labels. This set is used during the training process. The second set, Set_{Emo} , was collected using positive and negative emoticons as queries to the Twitter Search API. From the collected tweets, 500 positive and 500 negative were

manually annotated for each language by 2 volunteers. The third set Set_{RW} , released by Sascha Narr and Albayrak (2012), contains 739 positive and 488 negative manually annotated English tweets, and 159 positive and 160 negative French tweets.

To obtain more emotion-related patterns during the extraction phase, the psychological words (PW) were obtained from psychological categories in a text analysis application called Linguistic Inquiry and Word Count (LIWC) (James W. Pennebaker and Booth, 2007).

4.2 Experimental Result

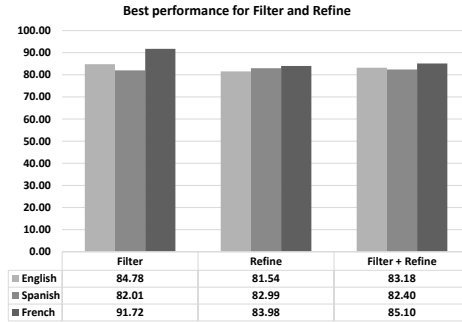


Figure 2: Filter + Refine performance with Set_{Emo} testing set

The first experiment evaluated the proposed approach using the Set_{Emo} testing set. The results in Figure 2 show that overall accuracies of over 80% can be obtained for all languages. Keep in mind that neither Filter nor Refine stages process all the posts, hence the reported individual accuracies are only over the portion of posts analyzed by each, with only the Filter+Refine result being over the totality of the posts. Due to time limitations, the approach used to combine the results from Filter and Refine is basic and treats both classifications as independent processes. A more elaborated approach may improve the results further.

Method	Lang	Patterns	Training Tweets	Accuracy
Narr	en	N.A.	≈ 500K	81.3 %
F+R	en	145	100K	79.0 %
Narr	fr	N.A.	≈ 100K	74.9 %
F+R	fr	100	50K	75.6 %

Table 1: Filter + Refine (F+R) compared to the related work (Narr).

Language	n-grams	patterns
English	141,981	145
Spanish	85,683	-
French	31,428	100

Table 2: Number of the extracted subsequences (n-grams), compared to the final number of patterns used in the comparison between Filter + Refine and Narr.

Finally, an experiment similar to the previous one was performed using Set_{RW} as the testing set. Table 1 shows that using a significantly lower number of features (the number of unigrams used by Narr is not available but must be large) and a smaller training set, the performance of the introduced approach surpasses the one reported by Narr for French and is just slightly lower for English. This is possible thanks to the effective reduction approach used to obtain relevant patterns from extracted subsequences, which is one of the main contributions of this paper (See Table 2). The training sizes used by Filter + Refine were limited due to data availability. More data would have probably helped Filter + Refine surpass the results reported by Narr for English.

5 Conclusion and Future Work

Two main types of language-independent features were studied in this paper: character-level n-grams, and emotion-bearing patterns. Character-level n-grams represent a useful tool for a preliminary sentiment classifier such as Filter. Emotion-bearing patterns can capture the emotional cues embedded in a person's writing. Such features can help identify subjectivity and determine the polarity of microblog posts across significantly different datasets in ways that regular patterns based purely on frequency wouldn't. It is believed that such emotion-bearing patterns can be used to perform a more complex analysis such as ambiguity and sarcasm identification, and to model other social and psychological characteristics of human behavior.

This paper contributes the introduction of emotion-bearing patterns as language independent features for multi-lingual sentiment analysis, and the efficient reduction approach used during their extraction. Sentiment analysis methods could benefit from such an approach during the training phases. Moreover, since the features obtained are very relevant to the classification task, less training examples are required, reducing the number of features significantly. Finally, the approach here presented is highly configurable, with both Filter and Refine relying on different thresholds. Several experiments with different sets of values can be performed to find the optimal set for a given language and show the full potential of the approach.

As a future work, it is planned to study the applicability of the presented approach to Asian languages such as Chinese and Vietnamese. Additionally, a deeper analysis of the patterns will be performed to extend the classification from the classic binary approach to a multi-class approach (using 8 different emotions). More difficult tasks such as detecting ambiguity and sarcasm will also be addressed.

References

- Anqi Cui, Min Zhang, Yiqun Liu, and Shaoping Ma. 2011. Emotion tokens: bridging the gap among multilingual twitter sentiment analysis. In *Proceedings of the 7th Asia conference on Information Retrieval Technology, AIRS'11*, pages 238–249, Berlin, Heidelberg. Springer-Verlag.
- Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 519–528, New York, NY, USA. ACM.
- Dmitry Davidov and Ari Rappoport. 2006. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, ACL-44*, pages 297–304, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters, COLING '10*, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Molly Ireland Amy Gonzales James W. Pennebaker, Cindy K. Chung and Roger J. Booth. 2007. The development and psychometric properties of liwc2007.
- Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of the International Conference on Computational Linguistics/Association*, pages 113–120, Sydney, Australia, 17th-21st July. ACL Press.
- Michael Hulphenhaus Sascha Narr and Sahin Albayrak. 2012. Language-independent twitter sentiment analysis. In *Workshop on Knowledge Discovery, Data Mining and Machine Learning*.
- Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods.
- Xuerui Wang, Andrew McCallum, and Xing Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM*, pages 697–702. IEEE Computer Society.
- Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. 2003. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*.