Textual and Contextual Patterns for Sentiment Analysis over Microblogs

Fotis Aisopos[$], George Papadakis[$,◇], Konstantinos Tserpes[$], Theodora Varvarigou[$]

[◇] L3S Research Center, Germany papadakis@L3S.de

[$] ICCS, National Technical University of Athens, Greece {fotais, gpapadis, tserpes, dora}@mail.ntua.gr

ABSTRACT

Microblog content poses serious challenges to the applicability of sentiment analysis, due to its inherent characteristics. We introduce a novel method relying on content-based and context-based features, guaranteeing high effectiveness and robustness in the settings we are considering. The evaluation of our methods over a large Twitter data set indicates significant improvements over the traditional techniques.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Information filtering; I.2.7 [**Artificial Intelligence**]: Natural Language Processing — Text analysis

General Terms

Algorithms, Experimentation

Keywords

Sentiment Analysis, N-gram Graphs, Social Context, Social Media

1. INTRODUCTION

The inherent characteristics of the text that is shared in social media infuse challenges in the extraction of sentiment-expressive patterns [4, 5, 6] that call for a different approach than those commonly followed by existing Sentiment Analysis (SA) systems. To this end, they employ either discriminative (series of) words [1] or dictionaries that assess the meaning and the lexical category of specific words and phrases (e.g. SentiWordNet[1]). Such characteristics are *sparsity* (short free-form text), *neologisms*[2],*noise* (misspelled text) and *multilinguality*. We introduce a novel approach based on two *complementary* sources of evidence, which are language-neutral and robust to noise: a content-based approach using the n-gram graphs document representation model and a context-based approach relying on social graph connections to capture the mood expressed in the *social context* of each message. We apply our approach in a large Twitter dataset and compare between these two sources of evidence and analytically examine how they perform in conjunction. We focus on effectiveness and efficiency, experimenting on multiple classification algorithms and configurations, aiming at the lower possible processing time.

[1] `http://sentiwordnet.isti.cnr.it`

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. PROBLEM FORMULATION

In this work, we exclusively focus on document-level SA in the context of microblog posts, detecting the sentiment "polarity" of individual Twitter messages (tweets). This can be categorized into two distinct problems:

- *Binary Polarity Classification*: Classify a document collection into two binary polarization classes $\mathcal{P}_{\mathcal{B}} = \{negative, positive\}$.
- *General Polarity Classification*: Classify a document collection into three polarization classes $\mathcal{P}_{\mathcal{G}} = \{negative, neutral, positive\}$.

3. APPROACH

3.1 Content-based Models

The alternative representation models used in the content-based approach, to extract the polarity features in each document (tweet) are the following:

- *Term Vector Model*: this model aggregates the set of distinct words contained in a document to represent as a vector of frequencies.
- *Punctuation Model*: this model takes into account the punctuation and character-based features that are contained in a document such as: (i) number of special characters, (ii) number of "!", (iii) number of quotes, (iv) number of "?", (v) number of capitalized tokens, (vi) length in characters.
- *Character N-grams Model*: this model comprises all substrings of length n in a document. This model constructs a vector providing the n-gram frequencies.
- *Character N-gram Graphs Model*: this model forms a graph whose nodes correspond to distinct n-grams, while its edges are weighted proportionally to the average distance - in terms of n-grams - between the adjacent nodes.

3.1.1 Character N-Gram Graphs Model

In the N-gram Graphs model, each polarity class is modeled by a single graph, uniformly aggregating the documents comprising it. After merging all individual document graphs into the class graph, its edges encapsulate the most characteristic patterns contained in the class' content, such as recurring and neighboring character sequences, special characters, and digits.

To estimate the similarity between a new document (i.e., tweet) graph G_{t_i} and a class graph G_{T_p}, we employ one of the established n-gram graph similarity metrics [3]:
(i) *Containment Similarity* (**CS**), which expresses the proportion of edges of a small graph G_{t_i} that are shared with graph G_{T_p}.

	Problem 1						Problem 2					
	4-Gram Graphs	Discr. Graphs	Punct.	Polarity	Discr. Polarity	Social Context	4-Gram Graphs	Discr. Graphs	Punct.	Polarity	Discr. Polarity	Social Context
NB	91.51%	96.36%	56.64%	53.40%	74.61%	51.05%	75.82%	93.43%	44.69%	37.40%	60.02%	34.33%
C4.5	**98.76%**	97.17%	60.98%	80.08%	72.89%	60.44%	**96.85%**	94.98%	46.00%	66.55%	61.47%	46.38%
SVM	86.10%	84.57%	50.12%	73.19%	72.89%	56.93%	79.18%	78.82%	39.02%	52.86%	57.27%	36.68%

Table 1: Accuracy of all combinations between models and classification algorithms over both polarity problems.

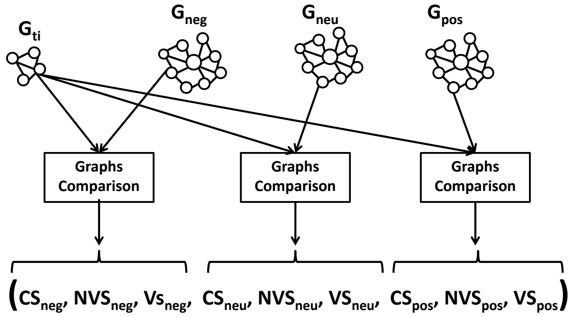
G_{ti}
G_{neg}
G_{neu}
G_{pos}
Graphs Comparison
Graphs Comparison
Graphs Comparison
$(CS_{neg}, NVS_{neg}, VS_{neg}, CS_{neu}, NVS_{neu}, VS_{neu}, CS_{pos}, NVS_{pos}, VS_{pos})$

Figure 1: Deriving the feature vector from the n-gram graphs model for General Polarity Classification Problem.

(ii) *Value Similarity* (**VS**), which indicates how many of the edges contained in graph G_{t_i} are shared with graph G_{T_p}, considering also their weights.
(iii) *Normalized Value Similarity* (**NVS**), which decouples value similarity from the effect of the largest graph's size.

In order to enhance the classification efficiency of the n-gram graphs model, we propose an intuitive method for discretizing its similarity values, employing pair-wise comparisons between the values of the same metric for different polarity classes, to produce a nominal label.

Figure 1 depicts the described process for estimating polarity between graph G_{t_i} and the three class graphs (G_{neg}, G_{pos}, G_{neu}) in the General Polarity Classification Problem.

3.2 Context-based Models

The 2 representation models for the context-based approach, aim at quantifying the effect of social context, along with their features.

3.2.1 Social Polarity Model

The aggregate sentiment of a set of tweets is determined by the dominant polarity class: if the positive messages significantly outnumber the negative ones, the overall sentiment is considered positive and vice versa. We consider the following context-based features:

- *Author Polarity Ratio*: the aggregate polarity of all messages posted by the same author
- *Author's Followees Polarity Ratio*: the aggregate sentiment of all messages posted by the author's followees
- *Author's Reciprocal Friends Polarity Ratio*: the aggregate sentiment of the tweets posted by the author's reciprocal friends
- *Topic(s) Polarity Ratio*: overall sentiment of all tweets that pertain to the same topic
- *Mention(s) Polarity Ratio*: the overall sentiment of all tweets that mention the same user
- *URL(s) Polarity Ratio*: aggregate polarity of all tweets with the same URL

3.2.2 Social Context Model

To reduce the feature extraction cost of the above model, we also consider an alternative set of context-based features that can be directly derived from a user's account and the characteristics of her messages. These features are the number of: Author's Tweets, Author's Followees, Author's Reciprocal Friends, Author's Reciprocal Friends, Topics, Mentions, URLs. These features rely on the same evidence with the Polarity Ratio model, but do not take into account the aggregate polarity of the underlying instances.

4. EVALUATION

Dataset. To examine the performance of our models, we conducted a thorough experimental study on a large-scale data set that was employed in [7]. To measure the effectiveness of the classification models, we considered the established metric of **classification accuracy** α.

Evaluation Method. To evaluate the performance of our models, we employ the 10-fold cross-validation approach. For the comparative analysis of the document representation models, we employed the Naive Bayes Multinomial (**NBM**) and the Support Vector Machines (**SVM**). For the rest of the models, we employed: Naive Bayes (**NB**), C4.5 and the SVM. For the functionality of the n-gram graphs, we employed the open source library of JInsect[2]. For the implementation of the classification algorithms, we used the Weka open source library[3].

Evaluation Results. Table 1 summarizes all experimental results over both polarity classification problems: On the whole, the four-gram graphs achieve the highest accuracy across all representation models and classification algorithms - especially after discretizing their values (as explained in Section 3.1.1). This means that they are more suitable for tackling the inherent characteristics of microblog content.

5. ACKNOWLEDGEMENT

This work has been partly funded by the FP7 EU Project SocIoS (Contract No. 257774).

6. REFERENCES

[1] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *COLING*, 2010.

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *EMNLP*, 2010.

[3] G. Giannakopoulos, V. Karkaletsis, G. A. Vouros, and P. Stamatopoulos. Summarization system evaluation revisited: N-gram graphs. *TSLP*, 5(3), 2008.

[4] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent Twitter sentiment classification. In *COLING*, 2011.

[5] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 2010.

[6] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, 2010.

[7] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, pages 177–186, 2011.

[2] http://sourceforge.net/projects/jinsect

[3] http://www.cs.waikato.ac.nz/ml/weka