Enhance Polarity Classification on Social Media through Sentiment-based Feature Expansion

Federico Alberto Pozzi
University of Milano-Bicocca
Viale Sarca, 336 - 20126
Milan, Italy
federico.pozzi@disco.unimib.it

Elisabetta Fersini
University of Milano-Bicocca
Viale Sarca, 336 - 20126
Milan, Italy
fersini@disco.unimib.it

Daniele Blanc
University of Milano-Bicocca
Viale Sarca, 336 - 20126
Milan, Italy
d.blanc@campus.unimib.it

Enza Messina
University of Milano-Bicocca
Viale Sarca, 336 - 20126
Milan, Italy
messina@disco.unimib.it

***Abstract*—Online social networking communities usually exhibit complex collective behaviors. Since emotions play a relevant role in human decision making, understanding how online networks drive human mood states become a task of considerable interest. One of the most relevant task in Sentiment Analysis is Polarity Classification, aimed at classifying the sentiment behind texts. We formulated different assumptions regarding which patterns within a message can be relevant sentiment indicators. Differently from well-formed texts, messages on social networks contain emoticons which could be strong sentiment indicators. For this, the first assumption states that the occurrences of emoticons representing a certain polarity could strongly agree with the overall message polarity. We then expanded the feature space including initialisms for emphatic and onomatopoeic expressions (e.g. bleh, wow, etc.) and "stretched words" (words with a letter repeated several times to emphasize a mood), extensively used in social media messages, because they could be useful information to help in determining the sentiment. Detailed analyses have been performed in order to support our assumptions. Four Machine Learning (supervised) classifiers are applied upon the expanded feature space model. Several experiments show that the considered features lead to increments of accuracy up to 5%.**

I. Introduction

According to the definition reported in [1], sentiment *"suggests a settled opinion reflective of one's feelings"*. The aim of Sentiment Analysis (SA) is therefore to define automatic tools able to extract subjective information, such as opinions and sentiments from texts in natural language, in order to create structured and actionable knowledge to be used by either a Decision Support System or a Decision Maker [2].

Sentiment Analysis is a growing area of Natural Language Processing with research ranging from document level classification [1] to learning the polarity of sentences [3] or features/objects [4]. The most widely studied problem is SA at document level [5], in which the naive assumption is that each document expresses an overall sentiment. When this is not ensured, a lower granularity level of SA could be more useful and informative.
Given the common characteristic of posts on social media to be short (e.g. the limit imposed by Twitter - a popular microblogging social networking web site - is 140 characters per post), classifying the sentiment of posts is most similar to sentence-level Sentiment Analysis.
However, the informal, specialized and length constrained language makes SA on social media a complex task. How well the features and techniques used on more well-formed data will transfer to the social media domain is an open question.

Characteristics that distinguish social media contents from well-formed contents (e.g. movie reviews [2], blogs or microblogs [6], and news [7]) is that review-type data often consists of relatively well-formed, coherent and at least paragraph-length pieces of text. Furthermore, resources such as polarity lexicons are usually available for these domains.
However, SA on social media leads towards new and more complex scenarios. In a post, a sentiment is conveyed in one or two sentence passages, which are rather informal and usually filled with abbreviations and typos. These messages are less consistent in terms of language, and usually cover a much wider array of topics. Since 2001, several studies based on polarity classification for well-formed scenarios have been proposed [4], [8], while polarity classification on user-generated content has rapidly grown only the last few years.
For instance, Barbosa and Feng [10] explored the linguistic characteristics of how tweets are written and the meta-information of words for polarity classification. In the study of Davidov et al. [11] four different feature types (punctuation, words, n-grams and patterns) are used for polarity classification and the contribution of each feature type evaluated for this task. Celikyilmaz et al. [12] proposed a new method for text normalization and investigated its effect when used for polarity classification. In particular, they used pronunciations of words to map alternative and shorter spellings into the intended words (reducing the sparseness caused by the noise in tweets).

SA is a multidisciplinary field that affects different branches of Computer Science, Social and Management Sciences. During the last years, several intersections between Sentiment Analysis and the Multi-Agent system technology are emerging [13]–[15]. For instance, Almashraee et al. [16] proposed a method that uses both multi-agent system technologies and machine learning techniques to provide a solution to the problems of polarity classification of on-line product features.They are able to extract data from several social media networks (one agent per network) and analyze sentiments using a learning mechanism for future predictions.
By using a stochastic multi-agent based approach, the system proposed by Gatti et al. [17] models and simulates user behavior on real-world social networks, taking into account what users (agents) post. Several challenges are faced: sampling the networks from the real-world social networks, performing text classication (Natural Language Processing) to predict topic

and sentiment from posts, modeling the user behavior to predict his/her actions (pattern recognition), and large-scale simulation.

In this work, we propose different approaches for text normalization and feature expansion to improve classification performance: after that messages are analyzed and normalized/preprocessed, different additional features (initialisms for emphatic and onomatopoeic expressions, emoticons, adjectives and "stretched words"[1]) are integrated within the bag-of-words model and used by common Machine Learning (supervised) classifiers. Further details are reported in Sect. II.

Although adaptation of this framework to other microblogs is straightforward, experiments are addressed on Twitter (tweets are short status updates of 140 characters or less) since it is an increasingly popular platform able to convey opinions and thoughts. Polarity classification approaches based on Twitter could provide unprecedented utility for different parties (e.g. marketing and financial purposes). For instance, an industry could gauge its recent marketing campaign by aggregating user opinions regarding their products. Moreover, it might be possible to identify the sentiment of financial news to forecast returns of markets [18] or sound out public opinion during political campaigns [19].

II. System architecture

We propose a system that is composed of three main modules: the first deals with preprocessing techniques, such as Text Normalization and Spelling Corrections, the second with Feature Expansion and the last with supervised classification techniques. Data are stored in a database to be subsequently easily reused and plotted. The system architecture is reported in Fig. 1.

A. Preprocessing module

Since tweets are similar to SMS messages, the writing style and the lexicon is widely varied. Moreover, tweets are often highly ungrammatical, and filled with spelling errors.

a) Text Normalization: In order to clean the dataset, we captured a set of patterns which are detected using dictionaries a priori defined and regular expressions. The applied filters are:

- **URLs**: All tokens matching the REGEXP `(https?|ftp|file)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|]` are transformed in its form without punctuation to avoid URL segmentation during tokenization (e.g. http://www.mind.disco.unimib.it becomes httpwwwminddiscounimibit);
- **Hashtags**: The symbol # is removed from all the tokens;
- **Mention Tags**: The tokens corresponding to a mention tag, identified through the REGEXP `@(.+?)`, are removed;
- **Retweet Symbols**: All the tokens matching the expression `RT @(.+?):` are removed.

[1]Words that have a letter repeated several times to emphasize a mood, e.g. "I'm so happyyyyyyyyyy!"

Note that the adaptation and modifications of the REGEXPs adopted in these filters to other microblogs is straightforward.

b) Spell-Checker: In addition to filters, misspelled tokens have been corrected using the Google's Spell Checker API[2]. Since the Google's algorithm takes the neighbourhood (context) of a misspelled token into account in suggesting the correction, the whole previously filtered tweet is considered as a query rather than the single token.

B. Feature Expansion module

Once the text normalization step has been performed, some additional features have to be extracted:

- **Emoticons**: in order to detect positive, neutral and negative emoticons, three dictionaries have been defined. For instance, positive emoticons are ':-)', ':)', '=)', ':D', neutral emoticons are ':-|', ':|', '=|', ';|' and negative emoticons are ':-(', ':(', '=(', ';('.
  If a token appears in the dictionary of positive emoticons then it is replaced with `POSEXPRESSIONS`, if it appears in the dictionary of neutral emoticons it is replaced with `NEUEXPRESSIONS`, otherwise with `NEGEXPRESSIONS`;
- **Initialisms for emphatic expressions**: several emphatic expressions are used in English. For instance, expressions such as 'ROFL','LMAOL','LMAO','LMAONF' represent positive expressions. They are replaced with `POSEXPRESSIONS`, `NEUEXPRESSIONS` or `NEGEXPRESSIONS`;
- **Slang correction**: in order to aggregate terms with the same meaning but different slangs, a dictionary of a priori defined slang expressions with their meaning, such as 'btw' (by the way), 'thx' (thanks), 'any1' (anyone) and u (you) has been built;
- **Onomatopoeic expressions**: as the previous point, a mapping dictionary has been defined for onomatopoeic expressions, such as 'bleh' (`NEGEXPRESSIONS`) and 'wow' (`POSEXPRESSIONS`). Also laughs are considered as onomatopoeic expressions: if a token has a sub-pattern matching `((a|e|i|o|u)h|h(a|e|i|o|u))\\1+|(ahha|ehhe|ihhi|ohho|uhhu)+`, then the whole token is replaced with `POSEXPRESSIONS`;
- **Stretched words**: a specific procedure has been defined to detect weather a term is a stretched word or not:

```
if Term has a lengthening then
    root ← Extract term root
    correction_list ← GoogleSpellChecker (root)
    if correction_list = ∅ then
        return isStretched
    end if
else
    return isNOTStretched
end if
```

[2]https://code.google.com/p/google-api-spelling-java/

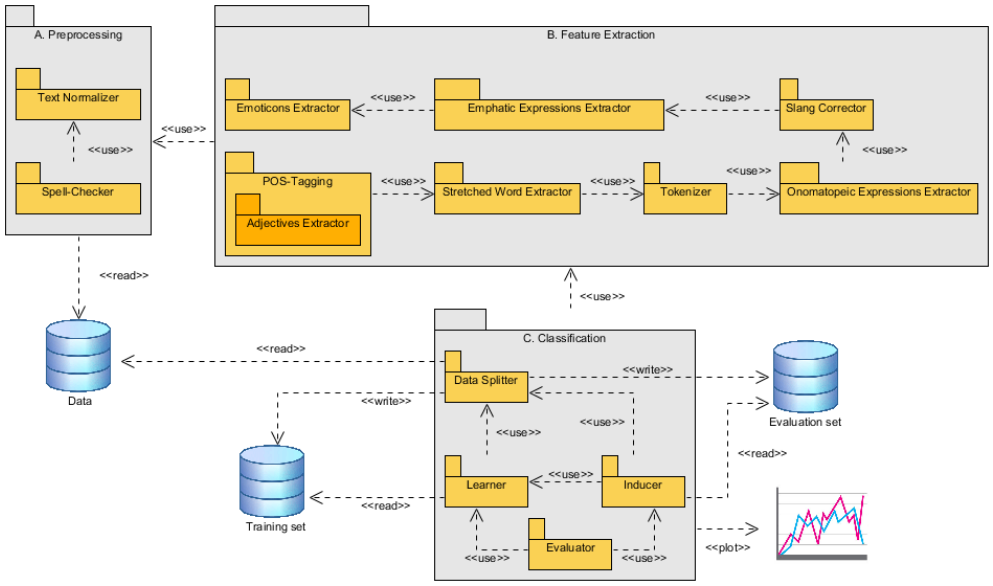


Fig. 1: System Architecture

A particular REGEXP has been defined to detect the presence of a lengthening (or *stretching*) in a term. Whether there is a match, the term root is extracted, otherwise the term is discarded and the next token is analyzed. The term root is analyzed by the Google's Spell Checker[3]. The spell checker's output is a list of possible corrections, ordered with respect to their probability. If the list is empty, the term is declared to be a stretch word.

- **Adjectives**: a Part-Of-Speech (POS) tagging process has been performed in order to tag each term with respect to its verbal form, to subsequently extract the adjectives (tagged as JJ, JJR, JJS) and determine their polarity depending on the fact that the term is in the dictionary of positive or negative terms. If the adjective is neither in the positive nor in the negative dictionary, its polarity is assumed to be neutral. The Stanford Log-linear Part-Of-Speech Tagger library[4] of Stanford University has been used for this task.

C. Classification module

Let $\vec{d} = (t_1, ..., t_n)$ a traditional feature vector composed only of terms, the new representing feature vector is defined as:

$$\vec{d}_{new} = (t_1, ..., t_n, e_{pos}, e_{neu}, e_{neg}, se, adj_{pos}, adj_{neu}, adj_{neg}, strw, class)$$

where $\{pos, neu, neg\} \in pol$ is the polarity, e_{pol} represents the emoticons, initialisms for emphatic and onomatopoeic expressions according to polarity pol, adj_{pol} represents the adjectives according to polarity pol, $strw$ represents the stretched words and $class$ is the ground truth polarity. According to the used term weighting method, boolean (0/1) or Term-Frequency (TF), adj_{pol} and $strw$ represents the presence or absence of the feature or how many times the feature occurs, respectively. Considering the boolean weighting schema, e_{pol} is zero if all the three atomic features involved (emoticons, initialisms for emphatic and onomatopoeic expressions) are zero and one if at least one of them exists, while considering the TF method it assumes the sum of how many times each atomic feature occurs. Experiments on the datasets have returned higher results using the 0/1 weighting schema for terms and TF for the additional external features.

The supervised classifiers used and compared in the system are: Naive Bayes (NB), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and Decision Trees (DT). Several classifier configurations have been tested (linear, polynomial and gaussian kernel for SVM and Naive Bayes Multinomial) and the most performing are used. For SVM, the linear kernel is used, while the Naive Bayes Multinomial overperforms the Naive Bayes classifier. K-NN has been tested for $k = 1, 3, 5, 10$ and the most performing value is $k = 3$. A 10-folds cross validation has been adopted as evaluation criteria. In order to obtain more statistically significant results, each experiment has been performed 10 times. The final performance are obtained by the arithmetic mean among the experiments. Classification experiments have been performed using Java. In particular, the WEKA[5] libraries have been adopted as tools for classification.

III. Datasets

In order to evaluate the proposed method, we performed our experiments on three datasets. The first is called **Gold Standard Person** [20], the second **Gold Standard Movie** [20] and the third is a concatenation of the two plus 3258 additional posts (that we called '**merged**').

Each gold standard dataset originally contains 1,500 manually labeled Twitter data for target-specific sentiment (i.e.,

[3]https://code.google.com/p/google-api-spelling-java/
[4]http://nlp.stanford.edu/software/tagger.shtml
[5]http://www.cs.waikato.ac.nz/ml/weka/

TABLE I: Precision, Recall and F-Measure per class on the Merged Dataset. The expression (v/ /*) indicates whether the values is statistically more (v), equally ( ) or less (*) significant (95% of confidence) than the baseline configuration (C).

Label	Classifier	Precision				Recall				F-Measure			
		C	C-F	NC-F	NC	**C**	C-F	NC-F	NC	**C**	C-F	NC-F	NC
POS	NB Multinomial	0,67	**0,7 v**	**0,7 v**	0,67	0,75	**0,79 v**	**0,79 v**	0,76	0,7	**0,74 v**	**0,74 v**	0,71
	SMO (PKI e=1)	0,71	**0,74 v**	0,73 v	0,71	0,69	**0,73 v**	**0,73 v**	0,69	0,7	**0,73 v**	**0,73 v**	0,7
	Ibk (k=3)	0,69	**0,74 v**	0,71	0,68	0,19	0,26 v	**0,27 v**	0,19	0,3	0,38 v	**0,39 v**	0,3
	J48	0,64	0,69 v	**0,7 v**	0,64	0,54	**0,64 v**	**0,64 v**	0,54	0,59	**0,66 v**	**0,66 v**	0,59
	(v/ /*)		(4/0/0)	(3/1/0)	(0/4/0)		(4/0/0)	(4/0/0)	(0/4/0)		(4/0/0)	(4/0/0)	(0/4/0)
NEU	NB Multinomial	0,84	0,86 v	**0,86 v**	0,84	0,88	0,88	0,88	0,88	0,86	**0,87 v**	**0,87 v**	0,86
	SMO (PK e=1)	0,8	0,81 v	**0,81 v**	0,8	**0,9**	**0,9**	**0,9**	0,89	0,85	0,85 v	0,85	0,84
	Ibk (k=3)	0,62	0,63 v	**0,64 v**	0,62	**0,98**	**0,98**	0,97 *	**0,98**	0,76	**0,77 v**	**0,77 v**	0,76
	J48	0,74	**0,76 v**	**0,76 v**	0,73	**0,88**	**0,88**	**0,88**	**0,87**	0,8	**0,82 v**	**0,82 v**	0,8
	(v/ /*)		(4/0/0)	(4/0/0)	(0/4/0)		(0/4/0)	(0/3/1)	(0/4/0)		(4/0/0)	(3/1/0)	(0/4/0)
NEG	NB Multinomial	0,76	**0,78**	0,77	0,75	0,42	**0,5 v**	**0,5 v**	0,42	0,54	**0,61 v**	**0,61 v**	0,53
	SMO (PK e=1)	0,66	**0,72 v**	0,69	0,65	0,34	**0,39 v**	0,38 v	0,35	0,45	**0,5 v**	0,49 v	0,45
	Ibk (k=3)	0	0,02	0,02	**0,06**	0	0	0	0	0	0	0	0
	J48	0,37	0,53 v	**0,55 v**	0,37	0,17	0,28 v	**0,29 v**	0,18	0,23	0,36 v	**0,38 v**	0,24
	(v/ /*)		(2/1/0)	(2/6/0)	(0/4/0)		(3/1/0)	(3/1/0)	(0/4/0)		(3/1/0)	(3/1/0)	(0/4/0)

the sentiment towards a specific target such as a movie or a person).
Each line of the data entry follows the format (id, topic, content, polarity), where 'id' is the id of the tweet, 'topic' is the name of the movie/person talked in the tweet, 'content' is tweet content and 'polarity' is the sentiment polarity about the topic expressed in the tweet, which can be 'pos' (positive), 'neg' (negative), 'neu' (neutral), or 'no sentiment' (not considered in this work). Polarity distributions for each of the studied dataset are reported in Figure 2.
We report in the following the descriptive analysis for adjectives, emoticons and stretched words. Similar statistics have been obtained (but omitted) for initialisms for emphatic and onomatopoeic expressions.

IV. Descriptive Analysis

The assumptions we formulated regard which patterns behind messages can be relevant sentiment indicators. Wilson et al. [21] shows that using emoticons for the learning phase of a classifier can lead to performance improvements. Moreover, Marchetti-Bowick and Chambers [22] present an approach that instead uses distant supervision (using emoticons as ground truth for labels) to train a classifier on a dataset of tweets, achieving higher performance in polarity classification. For this reason we argue that the occurrences of emoticons representing a certain polarity could strongly agree with the overall message polarity.
In addition to emoticons, we expect that also adjectives could be relevant sentiment indicators: in human interactions, the use of adjectives offers to the author the possibility to describe, in the best possible way, the own subjectivity within the discourse. Moghaddam and Popowich [23] demonstrated that the inclusion of adjectives as features in the bag-of-words model improves the classification accuracy.
Finally, we assume that also stretched words, extensively used in social media posts, could be useful information to help in determining the sentiment. To the best of our knowledge, no studies consider the combination of adjectives, initialisms for emphatic and onomatopoeic expressions, emoticons and stretched words as possible additional features.

In order to verify whether the proposed preprocessing techniques and generated features improve classification performance, four experiment configurations have been considered for each studied dataset (Table II).

TABLE II: Experiment configurations

Configuration	Text Normalization	Feature Expansion
Content (C)	✓	✗
PreprocessedContent (PC)	✓	✗
Content-FeatureExpansion (C-FE)	✗	✓
PC-FeatureExpansion (PC-FE)	✓	✓

A. Adjectives

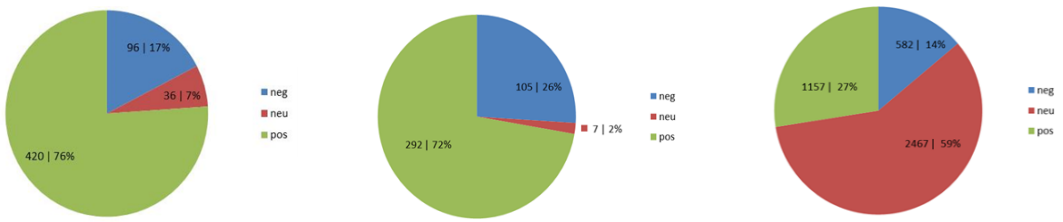
An analysis on the adjective distribution has been performed on the studied datasets (Table III).
First of all, as expected, positive and negative messages have a high percentage of adjectives. In order to verify that adjectives could be an important source of information for polarity classification, a further and detailed analysis has been conducted calculating conditional probabilities (conditioning the adjective presence to the overall message polarity) and with inverse conditional probabilities (viceversa). Conditional probabilities give us information about how much posts p classified with a contain polarity pol contain emoticons with the same polarity:

$$P(adj = pol \in p \mid p = pol) = \frac{\#(p = pol \wedge adj = pol \in p)}{\#(p = pol)}$$

where P stands for Probability.
Results (Table VI) generally show that the polarity of adjectives in messages agrees with the overall message polarity: this leads a positive message to have positive adjectives with a higher probability. Moreover, also inverse conditional probabilities have been calculated:

$$P(p = pol \mid adj = pol \in p) = \frac{\#(p = pol \wedge adj = pol \in p)}{\#(adj = pol \in p)}$$

as the ratio between the number of messages p with polarity pol that contain adjectives with polarity pol and the number of messages that contain adjectives with polarity pol.
The estimation of conditional probabilities and inverse conditional probabilities leads us to state the agreement between the adjective and message polarities.



(a) Gold Standard Movie Dataset (b) Gold Standard Person Dataset (c) Merged Dataset

Fig. 2: Polarity distribution

TABLE III: Adjective distribution

	Dataset	Tweets (%)	Polarity (%)	
With Adjectives	Movie	82% (455)	78% (353)	pos
			6% (29)	neu
			16% (73)	neg
	Person	67% (271)	76% (207)	pos
			2% (5)	neu
			22% (59)	neg
	Merged	54% (2262)	38% (850)	pos
			46% (1053)	neu
			16% (359)	neg
Without Adjectives	Movie	18% (97)	69% (67)	pos
			7% (7)	neu
			24% (23)	neg
	Person	33% (133)	64% (85)	pos
			2% (2)	neu
			35% (46)	neg
	Merged	46% (1944)	16% (307)	pos
			73% (1414)	neu
			11% (223)	neg

TABLE IV: Emoticon distribution

	Dataset	Tweets (%)	Polarity (%)	
With Emoticons	Movie	16% (88)	80% (70)	pos
			8% (7)	neu
			12% (11)	neg
	Person	7% (28)	14% (4)	pos
			0%	neu
			86% (24)	neg
	Merged	9% (381)	61% (231)	pos
			19% (74)	neu
			20% (76)	neg
Without Emoticons	Movie	86% (464)	75% (350)	pos
			6% (29)	neu
			18% (85)	neg
	Person	93% (376)	71% (268)	pos
			2% (7)	neu
			27% (24)	neg
	Merged	91% (3825)	24% (926)	pos
			63% (2393)	neu
			13% (506)	neg

B. *Emoticons*

Table IV shows the emoticons distribution on the three studied datasets. Positive and negative messages, as expected, have a high percentage of emoticons.
In order to verify that emoticons could be an important source of information for polarity classification (as well as adjectives), a further and detailed analysis has been conducted conditioning the emoticons presence to the message polarity and viceversa with inverse conditional probabilities. Conditional probabilities give us information about how much posts p classified with a contain polarity pol contain emoticons e with the same polarity:

$$P(e = pol \in p \mid p = pol) = \frac{\#(p = pol \wedge e = pol \in p)}{\#(p = pol)}$$

Results generally show that the polarity of emoticons in messages agrees with the message polarity: this leads a positive message to have positive emoticons with a higher probability. Moreover, as well as for adjectives, the inverse conditional probabilities have been calculated:

$$P(p = pol \mid e = pol \in p) = \frac{\#(p = pol \wedge e = pol \in p)}{\#(e = pol \in p)}$$

as the ratio between the number of messages p with polarity pol that contain emoticons e with polarity pol and the number of messages that contain emoticons with polarity pol.
Both probabilities further confirm the agreement between emoticons and message probabilities.

C. *Stretched words*

An further analysis has been performed on the stretched word distribution for the three studied datasets (Table V).
First of all, as expected, positive and negative messages have higher percentages of stretched words than neutral messages (even if messages which contain stretched words are very few).
In order to verify that stretched words could be an important source of information for polarity classification, a further and detailed analysis has been conducted conditioning the stretched words presence to the message polarity and viceversa with inverse conditional probabilities.
Conditional probabilities (and inverse conditional probabilities) are calculated as shown above for adjectives and emoticons. Supported from the analyzed data, we can conclude that stretched words have a high correspondence with positive and negative polarities.

TABLE V: Stretched words distribution

	Movie	Person	Merged
$P(stretch \in p)$	0,054	0,035	0,032
$P(stretch \notin p)$	0,946	0,965	0,968
$P(stretch \in p \mid p = pos)$	0,055	0,041	0,056
$P(stretch \in p \mid p = neu)$	0,028	0	0,013
$P(stretch \in p \mid p = neg)$	0,063	0,019	0,064

TABLE VI: Conditional probabilities for adjectives

(a) Movie

Positive Tweets		Neutral Tweets		Negative Tweets	
$P(adj \in p \mid p = pos)$	**0,840**	$P(adj \in p \mid p = neu)$	**0,806**	$P(adj \in p \mid p = neg)$	**0,760**
$P(adj \notin p \mid p = pos)$	0,160	$P(adj \notin p \mid p = neu)$	0,194	$P(adj \notin p \mid p = neg)$	0,240
$P(adj = pos \in p \mid p = pos)$	**0,643**	$P(adj = pos \in p \mid p = neu)$	0,278	$P(adj = pos \in p \mid p = neg)$	0,250
$P(adj = neu \in p \mid p = pos)$	0,386	$P(adj = neu \in p \mid p = neu)$	0,333	$P(adj = neu \in p \mid p = neg)$	0,375
$P(adj = neg \in p \mid p = pos)$	0,155	$P(adj = neg \in p \mid p = neu)$	**0,528**	$P(adj = neg \in p \mid p = neg)$	**0,448**

(b) Person

Positive Tweets		Neutral Tweets		Negative Tweets	
$P(adj \in p \mid p = pos)$	**0,709**	$P(adj \in p \mid p = neu)$	**0,714**	$P(adj \in p \mid p = neg)$	**0,562**
$P(adj \notin p \mid p = pos)$	0,291	$P(adj \notin p \mid p = neu)$	0,286	$P(adj \notin p \mid p = neg)$	0,438
$P(adj = pos \in p \mid p = pos)$	**0,524**	$P(adj = pos \in p \mid p = neu)$	0,143	$P(pos\,adj \in p \mid p = neg)$	0,124
$P(adj = neu \in p \mid p = pos)$	0,329	$P(adj = neu \in p \mid p = neu)$	**0,571**	$P(neu\,adj \in p \mid p = neg)$	**0,400**
$P(adj = neg \in p \mid p = pos)$	0,058	$P(adj = neg \in p \mid p = neu)$	0,143	$P(neg\,adj \in p \mid p = neg)$	0,238

(c) Merged

Positive Tweets		Neutral Tweets		Negative Tweets	
$P(adj \in p \mid p = pos)$	**0,735**	$P(adj \in p \mid p = neu)$	0,427	$P(adj \in p \mid p = neg)$	**0,617**
$P(adj \notin p \mid p = pos)$	0,265	$P(adj \notin p \mid p = neu)$	**0,573**	$P(adj \notin p \mid p = neg)$	0,383
$P(adj = pos \in p \mid p = pos)$	**0,528**	$P(adj = pos \in p \mid p = neu)$	0,120	$P(adj = pos \in p \mid p = neg)$	0,127
$P(adj = neu \in p \mid p = pos)$	0,373	$P(adj = neu \in p \mid p = neu)$	**0,327**	$P(adj = neu \in p \mid p = neg)$	**0,347**
$P(adj = neg \in p \mid p = pos)$	0,090	$P(adj = neg \in p \mid p = neu)$	0,042	$P(adj = neg \in p \mid p = neg)$	0,308

V. Results

In this section the performance achieved from the studied classifiers on the configurations of the Merged dataset are presented (Table I), since the Movie and Person datasets present few instances that could be not statistically significant (Figure 2). To this purpose, we measured Precision (P), Recall (R) and F_1-measure, defined as

$$P = \frac{TP}{TP+FP} \quad R = \frac{TP}{TP+FN} \quad F1 = \frac{2 \cdot P \cdot R}{P+R}$$

for the positive, neutral and negative labels. We also measured Accuracy, defined as

$$Acc = \frac{TP + TN}{TP + FP + FN + TN}$$

Figure 3 shows the final classification accuracy of each studied classifier on the Merged dataset[6]. The configurations C-FE and PC-FE lead to increments of approximately 2% of accuracy. In some cases, increments achieve 5%. Moreover, results of the two configurations are usually statistically significant (i.e. results are not randomly achieved). However, text normalization does not lead to significant improvements and results are not statistically significant. Regarding classifier performance, Multinominal Naive Bayes and SVM achieve the highest results.

VI. Conclusion

In this work, we proposed a system aimed at classify the polarity of messages on social media. We formulated different assumptions regarding what elements within a message can be relevant sentiment indicators. The first assumption states that the occurrences of emoticons representing a certain polarity could strongly agree with the overall message polarity. As well as expanding the feature space including emoticons, we assumed that also adjectives and stretched words, extensively used in social media messages, could be useful information to help in determining the sentiment. To the best of our knowledge, no studies consider the combination of adjectives, initialisms for emphatic and onomatopoeic expressions, emoticons and stretched words as possible additional features.
Subsequently, detailed analyses have been performed in order to verify our assumptions. For each studied dataset, four different configurations have been considered to measure the improvements led from each component (not preprocessed content and no additional features, not preprocessed content but additional features, preprocessed content but not additional features and preprocessed content with additional features).
The supervised classifiers used in the system are Naive Bayes (NB), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) and Decision Trees (DT). Several experiments show that text normalization does not lead to significant improvements but expanding the feature space of the traditional bag-of-words model with the considered features lead to accuracy increments up to 5%. Regarding classifier performance, Multinominal Naive Bayes and SVM achieve the highest results.

[6]Results on the other datasets are similar.

Acknowledgment

This work has been partially supported by the I-ShErPA project.

References

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1–135, 2008.

[2] F. A. Pozzi, E. Fersini, and E. Messina, "Bayesian model averaging and model selection for polarity classification," in *18th International Conference on Applications of Natural Language to Information Systems*, ser. LNCS. Springer Berlin Heidelberg, 2013, vol. 7934, pp. 189–200.

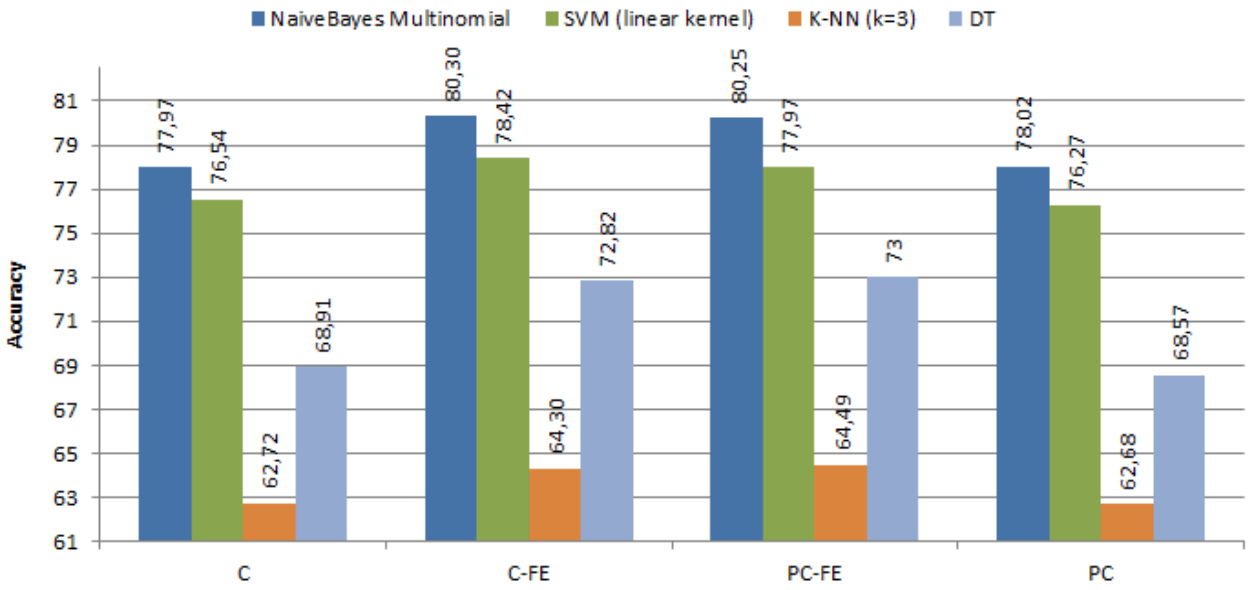


Fig. 3: Final results about the Merged dataset

[3] V. S. Jagtap and K. Pawar, "Analysis of different approaches to sentence-level sentiment classification," *International Journal of Scientific Engineering and Technology*, vol. 2, no. 3, pp. 164–170, 2013.

[4] H. Zhang, Z. Yu, M. Xu, and Y. Shi, "Feature-level sentiment analysis for chinese product reviews," in *3rd International Conference on Computer Research and Development (ICCRD)*, vol. 2, 2011, pp. 135–140.

[5] A. Yessenalina, Y. Yue, and C. Cardie, "Multi-level structured models for document-level sentiment classification," in *Proc. of the Conf. on Empirical Methods in NLP*, 2010.

[6] F. A. Pozzi, D. Maccagnola, E. Fersini, and E. Messina, "Enhance user-level sentiment analysis on microblogs with approval relations," in *AI*IA 2013*, ser. LNAI, M. B. et al., Ed. Springer International Publishing Switzerland, 2013, vol. 8249, pp. 133–144.

[7] Y. Rao, J. Lei, L. Wenyin, Q. Li, and M. Chen, "Building emotional dictionary for sentiment analysis of online news," *World Wide Web*, pp. 1–20, 2013.

[8] S. Mukherjee and P. Bhattacharyya, "Feature specific sentiment analysis for product reviews," in *13th International Conference on Intelligent Text Processing and Computational Linguistics*, ser. Lecture Notes in Computer Science, vol. 7181. Springer, 2012, pp. 475–487.

[9] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford, Technical Report, 2009.

[10] L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proc. of ACL*, 2010.

[11] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced sentiment learning using twitter hashtags and smileys," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, ser. COLING '10. Association for Computational Linguistics, 2010, pp. 241–249.

[12] A. Celikyilmaz, D. Hakkani-Tur, and J. Feng, "Probabilistic model-based sentiment analysis of twitter messages," in *Spoken Language Technology Workshop (SLT), IEEE*, 2010, pp. 79–84.

[13] D. Garcia and F. Schweitzer, "Emotions in product reviews-empirics and models," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on social computing (socialcom)*. IBaI Publishing, 2011, pp. 483–488.

[14] F. Schweitzer and D. Garcia, "An agent-based model of collective emotions in online communities," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 77, pp. 533–545, 2010.

[15] K. Fujimoto, "A computational account of potency differences in ewom messages involving subjective rank expressions," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on*, vol. 3, 2011, pp. 138–142.

[16] D. M. D. Mohammed Almashraee and R. Unland, "Sentiment classification of on-line products based on machine learning techniques and multi-agent systems technologies," in *Industrial Conference on Data Mining - Workshops*. IBaI Publishing, 2012, pp. 128–136.

[17] M. Gatti, A. P. Appel, C. Pinhanez, C. dos Santos, D. Gribel, P. Cavalin, and S. B. Neto, "Large-scale multi-agent-based modeling and simulation of microblogging-based online social network," in *The 14th International Workshop on Multi-Agent-based Simulation (MABS, AAMAS)*, 2013.

[18] G. Mitra and L. Mitra, *The Handbook of News Analytics in Finance*. John Wiley & Sons, Ltd., 2011.

[19] B. O'connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls : Linking text sentiment to public opinion time series," in *International AAAI Conference on Weblogs and Social Media*, 2010.

[20] M. N. S. W. Lu Chen, Wenbo Wang and A. P. Sheth, "Extracting diverse sentiment expressions with target-dependent polarity from twitter," in *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

[21] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Association for Computational Linguistics, 2005, pp. 347–354.

[22] M. Marchetti-Bowick and N. Chambers, "Learning for microblogs with distant supervision: political forecasting with twitter," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '12. Association for Computational Linguistics, 2012, pp. 603–612.

[23] S. Moghaddam and F. Popowich, "Opinion polarity identification through adjectives," *CoRR*, 2010.