SentiCorr: Multilingual Sentiment Analysis of Personal Correspondence

Erik Tromp, Mykola Pechenizkiy
Department of Computer Science
Eindhoven University of Technology
Eindhoven, P.O. Box 513, 5600MB,The Netherlands,
Email: m.pechenizkiy@tue.nl

***Abstract*—We present the system for automated sentiment analysis on multilingual user generated content from various social media and e-mails. One of the main goals of the system is to make people aware how much positive and negative content they read and write. The output is summarized into a database allowing for basic OLAP style exploration of the data across basic dimensions including for example time and correspondents dimensions. The sentiment analysis is based on a four-step approach including language identification for short texts, part-of-speech tagging, subjectivity detection and polarity detection techniques. We extensively tested our system on data from Twitter, Facebook and Hyves. We also developed an MS Outlook sentiment analysis plug-in allowing people to see how positive or negative the content of the e-mails is and provide confirmatory or correcting feedback on the correctness of the sentiment classification at the sentence or e-mail level.**

I. Introduction

We present SentiCorr – the system for automated sentiment analysis on multilingual user generated content from various social media and e-mails. One of the main goals of the system is to make people aware how much positive and negative content they read and write. The output is summarized into a database allowing for basic OLAP style exploration of the data across basic dimensions including time, correspondents, read/write and alike. We developed also MS Outlook plug-in that analyses the correspondence and highlights positive and negative sentences.

The sentiment analysis is based on a four-step approach including language identification, part-of-speech tagging, subjectivity detection and polarity detection. We developed LIGA, the Graph-based Approach for Language Identification presented in [7]. LIGA is tailored for short texts. For part-of-speech (POS) tagging we use an existing solution called the TreeTagger [6]. We apply AdaBoost [3] using decision stumps for subjectivity detection. For polarity detection we developed RBEM - a heuristic Rule-Based algorithm to create an Emissive Model on patterns.

We extensively tested our system on publicly available social media including Twitter, Facebook and Hyves – the largest social network in the Netherlands (http://www.hyves.nl/). The generalization performance of the sentiment analysis at the sentence level is close to 70% for the three-class classification *positive*, *negative*, *objective* on a balanced test set. The performance on the e-mail correspondence still awaits a larger scale evaluation for which we currently attract volunteers, willing to use the plug-in, and to provide relevance feedback and ideally the content of wrongly processed cases for further offline and online fine-tuning of the system.

The rest of the paper is organized as follows. In Section II, we concisely present the four step approach for sentiment classification and in Section III we give an insight on the system from the software perspective. Section IV concludes and discusses directions for further work.

II. Sentiment classification approach

Sentiment analysis can be performed at different levels of granularity with different levels of detail. We perform it at the sentence level. The level of detail typically goes into determining the *polarity* of a message, which is what we investigate as well. A more detailed approach could be to determine the emotion expressed [5] in addition to the polarity. This is not developed in the current system.

The problem we solve is as follows: taking a message as input we need to produces a polarity indication as output. The message originates from one of the social media sources (Twitter, Facebook and Hyves) or MS Outlook e-mails. The language of the message is not known beforehand and we are interested in messages written in a closed set of multiple languages rather than only regarding one language, for example English. It is not uncommon for a person to write and read correspondence in more than one language, e.g. in mother tongue(s), in English and in the language of the country the persons lives in.

Conceptually, we can regard the problem of sentiment analysis as shown in Figure 1 where we have a stream of unstructured texts originating from social media or e-mail server/client as input.

Determining the opinion of a message can be as extensive as regarding *objective* (not expressing any sentiment), *positive* (expressing positive sentiment), *negative* (expressing negative sentiment), *neutral* (expressing sentiment but neither positive nor negative) and *bipolar* (expressing both positive and negative sentiment) messages. SentiCorr however assumes that messages that are neutral are also objective. Moreover, as the messages placed on social media are very short, we also assume that only one polarity is prevailingly expressed in a message, we thus do not regard the bipolar class as we assume

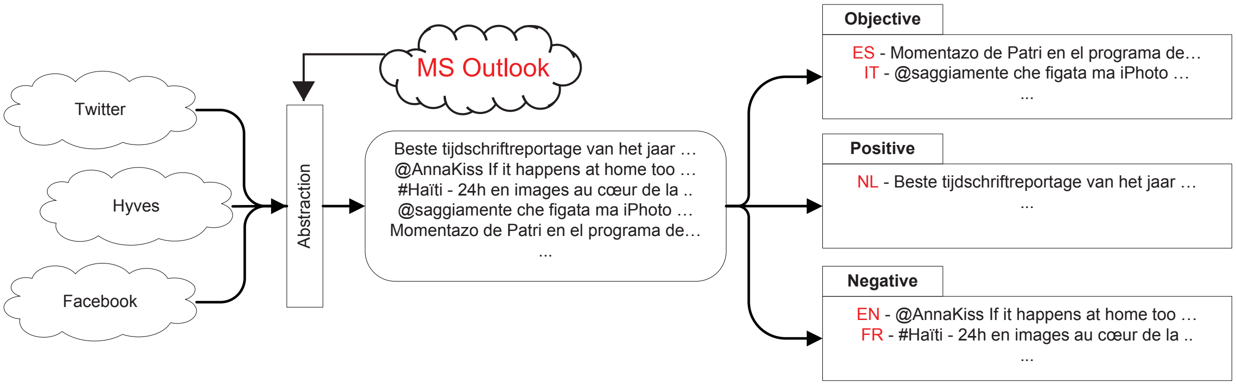

MS Outlook
Twitter
Hyves
Facebook
Abstraction
Beste tijdschriftreportage van het jaar …
@AnnaKiss If it happens at home too …
#Haïti - 24h en images au cœur de la ..
@saggiamente che figata ma iPhoto …
Momentazo de Patri en el programa de…
…
Objective
ES - Momentazo de Patri en el programa de…
IT - @saggiamente che figata ma iPhoto …
…
Positive
NL - Beste tijdschriftreportage van het jaar …
…
Negative
EN - @AnnaKiss If it happens at home too …
FR - #Haïti - 24h en images au cœur de la ..
…

Fig. 1. The conceptual idea of sentiment analysis. The input is a stream of social media and e-mail messages in different languages. The output is a crisp classification into objective, positive or negative along with the language of a message (and language of each phrase of an e-mail).

that a message containing two polarities always favors one of the two. For e-mails we perform the classification at phrase level, thus allowing an e-mail as a whole to be bipolar.

Additionally, we associate the language a text is written in with each text, allowing for more fine-grained sentiment analysis by exploiting knowledge on the language a message is written in. Knowing the language of each message also allows for segmenting based on the language dimension. The results of the sentiment analysis are aggregated in different ways to obtain high-level statistics and make it available for the user of the system.

Concisely stated, the answer to the problem we solve is a hard label being one of *objective*, *positive* or *negative*. The formulation of the problem itself is as follows: *Given a message written in an unknown language to be determined, does this message contain sentiment and if so, what is the polarity of this sentiment?*

The approach we take to solve the sentiment analysis problem consists of four steps; *language identification*, *part-of-speech tagging*, *subjectivity detection* and *polarity detection*, as shown in Figure 2. This implies that only subjective texts are processed by the polarity detection step as objective texts do not have any polarity. The rationale behind using four steps is that we can more specifically construct models in a later step when knowledge from earlier steps is present. The separation of subjectivity and polarity detection is inspired by [4].

For language identification we use a recently introduced algorithm called LIGA [7]. This algorithm uses a graph formalism to incorporate elements of the grammar into its model. Language identification is formulated as a supervised learning task. Given some historical data in which for each message we know a label, the language in which this text is written, our goal is to learn a model such that given some previously unseen message we can say as accurately as possible which language this message is written in. Thus, we learn a graph model on labeled data. The labels of graphs vertices represent the presence of words in a given language. The weights of the vertices represent the frequencies of words in a given language. The crucial part is in the presence and weights of the edges, which try to capture the ordering of N-grams (at the character level) in each of the languages.

For part of speech tagging (POS-tagging) we use existing work TreeTagger [6] having publicly available models for different languages http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. The main goal of this step is to expand the space of predictive features based on the knowledge of the certain language (POS-tags differ per language).

For subjectivity detection we use AdaBoost ensemble learning [3] with decision stumps as weak learners. The goal of subjectivity detection is to determine whether a given message is subjective or objective based on the training set containing labeled instances. By subjective we mean a message containing any form of private state; a general term covering opinions, beliefs, thoughts, feelings, emotions, goals, evaluations and judgements.

Finally, for polarity detection we introduced the Rule-Based Emission Model (RBEM) algorithm. Each entity in a message can emit positive or negative sentiment. The rules are defined on eight different types of patterns – positive, negative, amplifier, attenuator, right/left flip, continuator and stop patterns. The idea of using patterns arises from [8]. The emission aspect of our RBEM algorithm is related to smoothing which is often applied in different machine learning settings.

III. System perspective

Our SentiCorr system is implemented as an extendable framework presented in Figure 3. This framework allows to crawl and scrape social media for data that can serve as input for the sentiment analysis process. The analysis of the e-mails is implemented as a subsystem separated from the social media sources. This is done for the pragmatic reasons – people who use mainly e-mail or mainly social media can use one of the plug-ins. As the information retrieved from various social media and e-mails differs in format we use an abstraction layer such that all input looks the same to our sentiment analysis process.

The sentiment analysis itself in our setting consists of the four steps but due to the nature of the framework, more steps can be added without jeopardizing the other steps.

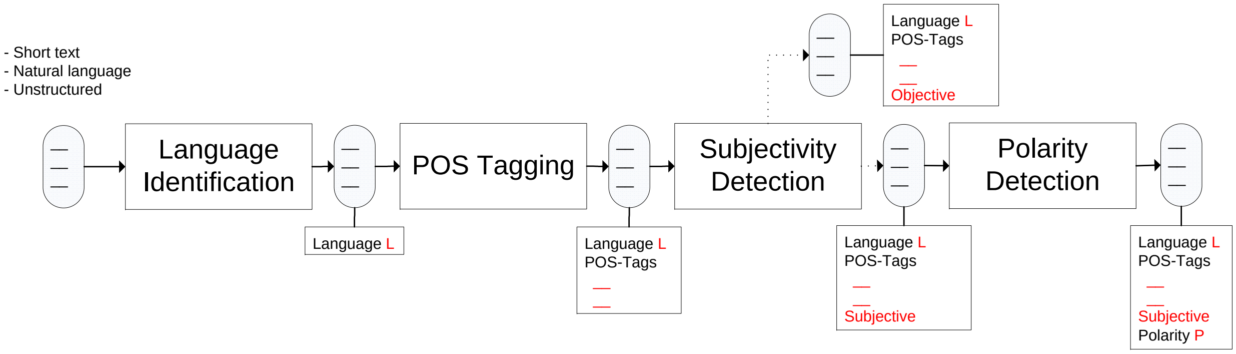


Fig. 2. The four-step approach for sentiment analysis. The input consists of unstructured texts. First, language identification determines the language after which POS-tagging enriches the text with additional features (POS-tags). The subjectivity detection determines if the text contains subjective aspects. If not, the text is objective, otherwise the text's polarity is determined by the polarity detection.

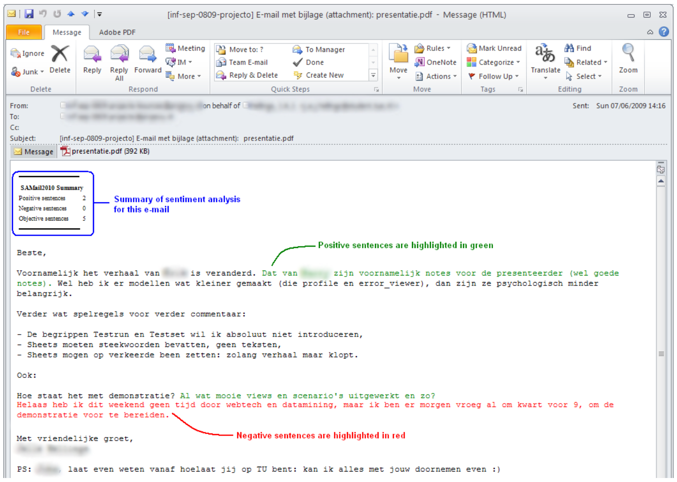
For example, an emotion classifier – a more elaborate case of sentiment analysis where the emotion of a message is automatically extracted – can be added as an additional step. Each of the separate steps of our framework can also be extended, modified or replaced to suit different purposes. As the output of one step is the input of the other step, there is no dependency between the actual operational properties of the steps but only on the input and the output. This means that any step can be replaced or extended in any way as long as the input and output satisfy the interface requirements. For example, it is possible to replace the current for subjectivity detection with Naive Bayes or other technique performing a two-way classification into subjective or objective.

Figure 4 illustrates one of the ways the information is presented to the user in an integrated view. Social media content and the summary views on the quantities of positive and negative content read or written over different periods of time are presented via separate interfaces providing an OLAP-style exploration of the data along the predefined dimension and allowing to zoom in to the level of the individual messages and zoom out to the grand total summary of the sentiments (see Figure 4 for an intuitive example).

IV. Conclusions and Future Work

We described SentiCorr – the system for automated sentiment analysis on multilingual user generated content from various social media and e-mails. Unlike most of the existing systems focusing on addressing marketing questions with sentiment classification and opinion mining functionality, our system is aimed to help individual users to become more aware of the sentiment in their correspondence.

The prospective use of SentiCorr: The current prototype can be used as is for monitoring the statistics collected from the sentiment analysis of the various texts and their further exploration. At the moment we have full support for English and Dutch languages in the developed four step sentiment analysis process. LIGA, the language identification part at the moment supports six European languages.



[inf-sep-0809-projecto] E-mail met bijlage (attachment): presentatie.pdf - Message (HTML)

From: [redacted] on behalf of [redacted]
Sent: Sun 07/06/2009 14:16
To: [redacted]
Cc:
Subject: [inf-sep-0809-projecto] E-mail met bijlage (attachment): presentatie.pdf

Message presentatie.pdf (392 KB)

SAMail2010 Summary	
Positive sentences	2
Negative sentences	0
Objective sentences	5

```
Beste,

Voornamelijk het verhaal van [redacted] is veranderd. Dat van [redacted] zijn voornamelijk notes voor de presenteerder (wel goede
notes). Wel heb ik er modellen wat kleiner gemaakt (die profile en error_viewer), dan zijn ze psychologisch minder
belangrijk.

Verder wat spelregels voor verder commentaar:

- De begrippen Testrun en Testset wil ik absoluut niet introduceren,
- Sheets moeten steekwoorden bevatten, geen teksten,
- Sheets mogen op verkeerde been zetten: zolang verhaal maar klopt.

Ook:

Hoe staat het met demonstratie? Al wat mooie views en scenario's uitgewerkt en zo?
Helaas heb ik dit weekend geen tijd door webtech en datamining, maar ik ben er morgen vroeg al om kwart voor 9, om de
demonstratie voor te bereiden.

Met vriendelijke groet,
[redacted]

PS: [redacted] laat even weten vanaf hoelaat jij op TU bent: kan ik alles met jouw doornemen even :)
```

Fig. 4. Highlighting of positive and negative sentences in the message body and providing a summary for each e-mail. At the moment the relevance feedback mechanism allowing a user to confirm correctly classified sentiments and correct wrongly classified sentiments is available as a separate tool.

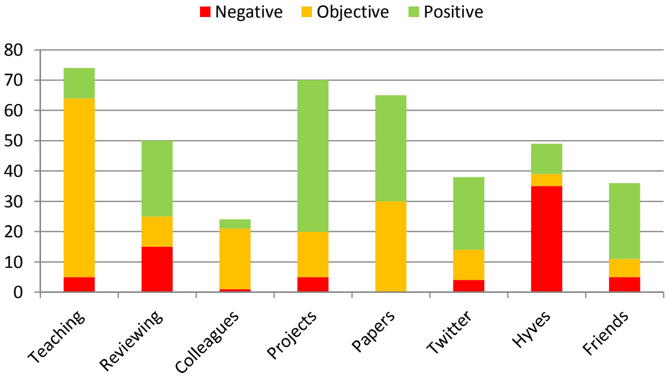


Fig. 5. A separate GUI interface allows basic exploration of the sentiment summaries along predefined dimensions and categories within each of the dimensions. E.g. a weekly summary of total (sent and received) positive, objective and negative content.

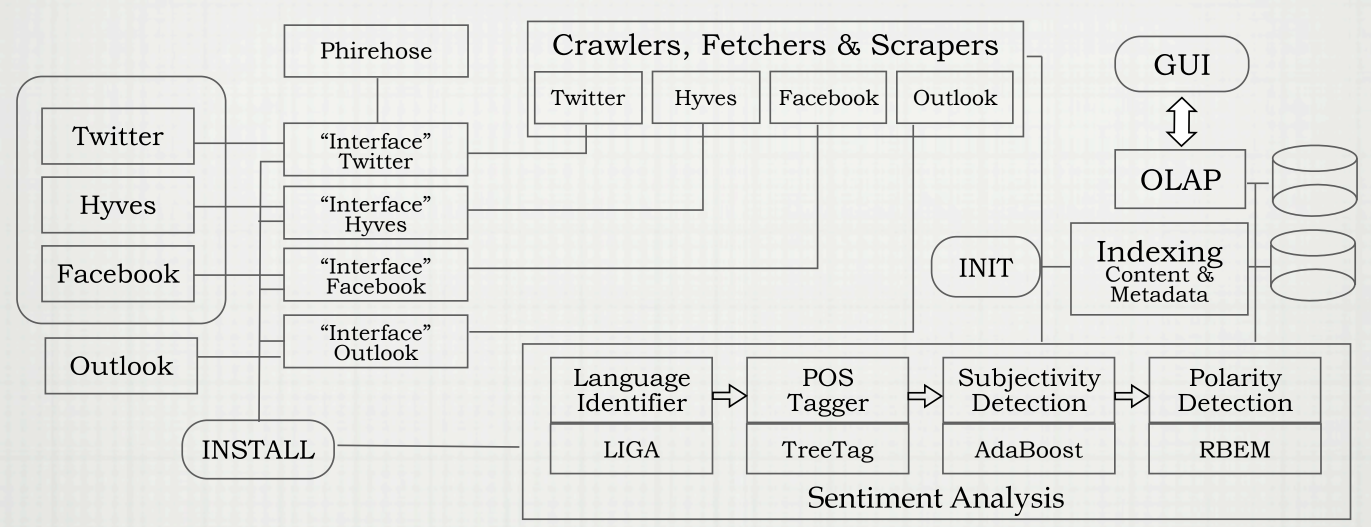


Fig. 3. A high level overview of the SentiCorr software framework.

Besides the stand-alone use of the developed system, we are interested in integrating SentiCorr into the so-called Stress Analytics system by relating the identified sentiments and their summaries with other events potentially related to the occurrences of stress. Information on such related events can be extracted from a stress measuring device (detecting arousal based on heart rate and galvanic skin response measurements), calendar data or working agendas, and analysis of speech and facial expressions. These data sources are known to be useful for categorizing the level of stress a person experiences. To give more context on the other prospective use of the developed software as an integral component of a larger system, we refer an interested reader to [1] and [2] that introduce the problem of measuring, understanding and managing stress at work. One very important step in the process of stress management is making the worker aware of the past, current or expected stress. SentiCorr thus will provide one kind of the valuable input for a stress analytics system.

Future work. We continue the development of the described system in several directions.

Obviously, we want SentiCorr to be as accurate as possible in the sentiment classification. In this context we find important to further improve user experiences in interacting with the system and to find ways encouraging the users to provide relevance feedback to the system.

Current approaches for sentiment classification focus on distinguishing positive sentiment from negative and those that consider neutral sentiment do not distinguish it from objectiveness. However, in many practical situations we are interested not only in the sentiment, i.e. attitude of a person, but the positiveness or negativeness of the information itself even if it presented in an objective way. For example, news in media or business news are considered to be objective, that is sentiment neutral. However, even being objective, news often contain positive or negative information, e.g. bad news about a nature hazard leading to financial, environmental and/or human losses are objective, but may seriously effect the person reading them. Similarly, personal messages positioned neutrally may contain very negative information for the person, e.g. "We have to fire every third employee within next 5 months because of the serious budget cuts." We set a goal to develop a new technique for determining the positiveness of the information in the objective messages.

Further information about the SentiCorr development will be made available at http://www.win.tue.nl/~mpechen/projects/senticorr/.

REFERENCES

[1] J. Bakker, L. Holenderski, R. Kocielnik, M. Pechenizkiy, and N. Sidorova. Stess@work: From measuring stress to its understanding, prediction and handling with personalized coaching. In *Proc. of ACM SIGHIT International Health Informatics Symposium (IHI 2012)*. ACM Press.

[2] J. Bakker, M. Pechenizkiy, and N. Sidorova. What's your current stress level? detection of stress patterns from gsr sensor data. In *Proc. of 2nd HaCDAIS Workshop collocated with IEEE ICDM 2011*. IEEE Press.

[3] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting, 1997.

[4] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004.

[5] D. Potena and C. Diamantini. Mining opinions on the basis of their affectivity. In *2010 International Symposium on Collaborative Technologies and Systems (CTS)*, pages 245–254, 2010.

[6] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proc. of the International Conference on New Methods in Language Processing*, pages 44–49, 1994.

[7] E. Tromp and M. Pechenizkiy. Graph-based n-gram language identification on short texts. In *Proceedings of the 20th Machine Learning conference of Belgium and The Netherlands*, 2011.

[8] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.