# Domain identification through sentiment analysis

Ricardo Martins, José João Almeida, Pedro Henriques, Paulo Novais
*Algoritmi Centre / Department of Informatics*
*University of Minho, Braga - Portugal*
*ricardo.martins@algoritmi.uminho.pt, {jj, prh, pjon}@di.uminho.pt*

Keywords: Emotional Profile, Sentiment Analysis, Machine Learning, Natural Processing Language

Abstract: When dealing with chatbots, domain identification is an important feature to adapt the interactions between user and computer in order to increase the reliability of the communication and, consequently, the audience and decrease its rejection avoiding misunderstandings. In order to adapt to different domains, the writing style will be different for the same author. For example, the same person in the role of a student writes to his professor in a different style than he does for his brother. This article presents a process that uses sentiment analysis to identify the average emotional profile of the communication scenario where the conversation is done. Using Natural Language Processing and Machine Learning techniques, it was possible to obtain an index of 96.21% of correct classifications in the identification of where these communications have occurred only analysing the emotional profile of these texts.

## 1 Introduction

Along the day, a person must represent different roles: worker, father, student, boss, ... and for each role he must interact to other people according to the place they are and the reaction or feedback he receives from others. In some cases it is necessary to be more "politically correct" during the speeches, and in other cases not so much.

According to Collins dictionary [3], "if you say that someone is politically correct, you mean that they are extremely careful not to offend or upset any group of people in society who have a disadvantage, or who have been treated differently because of their sex, race, or disability." When talking to people, in a daily interaction or via chatbots, the idea is to decrease the chances of being rejected by the target audience, increasing the chances of get his speech accepted. So, identifying the audience's communication profile is the first step to provide a better experience between users and chatbots. Knowing where the conversation takes place is essential to avoid misunderstandings. For example, like in the real life, it is unacceptable to talk to professors in classroom like we talk to best friends during a party. So, it is not acceptable that a chatbot responses to a user different than the pattern from where the conversation takes place.

The purpose of this article is to present a classifier based in sentiment analysis which compares speeches from public and common person in social media as LinkedIn and Twitter in order to identify the emotional communication profile for each social media and predict the domain where the conversation is being held. For "domain", we consider as the social media where the text was originally posted.

The remainder of this paper is as follows: Section 2, introduces the concept of emotion and presents the basic emotions theory for emotion representation and analysis. Section 3 presents some work in this area to detect emotion from social media, while Section 4, describes the steps followed in our emotional analysis in chatbot messages and discusses some results obtained, and finally, the paper ends in Section 5 with the conclusion and future work.

## 2 Basic emotions

Basic emotion theorists agree that all human emotion can be contained within a small set of basic emotions, which are discrete.

Many researchers have attempted to identify a number of universal basic emotions which are common for all people and differ one from another in important ways. A popular example is a cross-cultural study of 1972 by Paul Ekman [4] and his colleagues, in which they concluded that the six basic emotions are *anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*.

A major part of work in emotion mining and classification from text has adopted this basic emotion set. For

example, in order to model public mood and emotion, Bollen et al [2] extracted six dimensions of mood including *tension*, *depression*, *anger*, *vigour*, *fatigue*, *confusion* from Twitter. Strapparava and Mihalcea [11] created a large data set with six basic emotions: *anger*, *disgust*, *fear*, *joy*, *sadness* and *surprise*. For Plutchik [9], all sentiment is composed of a set of 8 basic emotions: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*.

However, there is no consensus on which human emotions should be categorized as basic and be included in the basic emotion set. Moreover, the emotional disambiguation is a contested issue in emotion research. For instance, it is unclear if *surprise* should be considered an emotion since it can assume negative, neutral or positive valence.

In our tests, it was used the Plutchik model because is a well-known model implemented in some libraries and toolkits for sentiment analysis used in this work.

# 3 Related work

Despite the vast amount of works using sentiment analysis, none of them considers the messages emotional profile as a dimension of the communication profile. So, each work cited below has inspired partially our work as will be mentioned.

Analysing emotions in social media was suggested by Schwartz et al [10], whose work predicts the individual well-being, as measured by a life satisfaction scale, through the language people used on social media communication channels. This is made using randomly selected posts from Facebook and a lexicon-based approach to identify the text words polarities.

Other work who inspired our analysis was introduced by Baldoni et al [1] who has presented a project involving lexicons and ontologies to extract emotions including sadness, happiness, surprise, fear and anger, which contributed in the emotional profile creation.

The work of Widmer [12] contributed with the idea of domain identification using machine learning techniques.

# 4 Data Analysis

In our analysis, all data was collected from same author's public posts and texts in LinkedIn[1] and Twitter[2]. The choice of these social media is because while the audience of LinkedIn is more professional and aimed at laboral relationships - and for these reasons more politically correct - Twitter has a different audience profile, aimed at casual relationships, i.e., people on Twitter tend to expose their opinions with more freedom.

The authors, as presented in Table 1, were selected randomly according to the following criteria:

- Must have LinkedIn and Twitter profiles;
- Must have at least 10 opinion texts published in LinkedIn pulse;
- Must have at least 500 posts in Tweeter.

So, having in mind these requirements and after a search at LinkedIn and Tweeter profiles, the authors mentioned in Table 1 were chosen to provide the texts for analysis.

Table 1: Authors analysed

| **Author** | **Profession** |
|---|---|
| C. Fairchild | News editor |
| J. Saper | Investor |
| J. Battelle[3] | Entrepreneur |
| B. McGovan | Media Trainer |
| A. Mitchell | Professor |
| L. Profeta | Medical Doctor |

## 4.1 Data preprocessing

Preprocessing is a data mining technique that transforms raw data into an understandable form. There are in the literature several preprocessing techniques available to extract information from text, and their usage is according to the characteristics of the information desired.

In order to analyse the emotion contained into the text, all texts have been preprocessed according to the planned pipeline described in Figure1, where each process is denoted by an acronym as follow:

1. TK - Tokenization;
2. POS-T - Part of Speech Tagging;
3. NER - Name Entity Recognition;
4. SWR - Stopwords Removal.

POS-T process identifies the text grammatical structure and tags all nouns, verbs, adverbs and adjectives removing the remaining words. The reason for this text cleaning is because only these grammatical categories can bring emotional information. In a formal description, the TK process converts the original text $D$ in a set of tokens $T = \{t_1, t_2, ..., t_n\}$ where each

[1] http://www.linkedin.com
[2] http://www.twitter.com

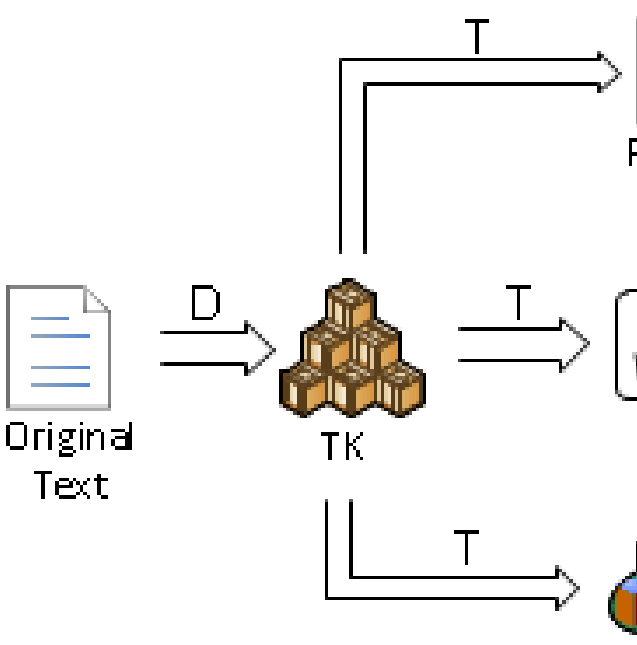


Figure 1: Preprocessing tasks

element contained in $T$ is part of the original document D. Later, the POS-T labels each token with a semantic information and creates a set P, where $P_T = \{p_{(T,1)}, p_{(T,2)}, ..., p_{(T,k)}\}$ and $0 \leq k \leq n$ and $P_T \subset T$, and P is-a noun, verb, adverb, adjective.

NER process separates (nouns) names in 3 different categories: "Location", "Person" and "Organization" and removes all tokens related with these categories. As result, a set $N_T = \{n_{(T,1)}, n_{(T,2)}, ..., n_{(T,j)}\}$ is constructed based on identified word category and where $0 \leq j \leq n$ and $N_T \subset T$. This step is important to be done in parallel with POS-T because some locations can be confused with some grammatical structure (as Long Beach or Crystal Lake, for instance).

SWR process is responsible of the removal of stopwords (undesirable words in the text) from the tokens. The stopwords gathered in a predefined set $SW = \{sw_1, sw_2, ...sw_y\}$ of words, available in R through the package **tm**[5] and the SWR process result is a set $T' = T - SW$.

After the 3 preprocessing tasks finish, the result document $PR$ must contain a set of words where $PR = T' \cap P_T \cap N_T$.

For all three tasks - POS-T, NER and TK - the Stanford Core NLP [7] toolkit was used.

## 4.2 Emotional analysis

After the preprocessing, all texts were analysed against EmoLex lexicon [8] in order to detect the amount of each basic emotion for each phrase according the Plutchik model. For this step, it was used the package Syuzhet [6] available in R.

The texts' emotional values returned by the Syuzet package is a scalar value, so it was necessary to convert them in a scale ranging from 0 to 1. This is important for determining the percentage of each basic emotion that is commonly used by the author in his texts. For this conversion, it was considered the ratio of each basic emotion in the emotion total amount. For example, the ratio (P) of the emotion (E) *fear* is calculated as:

$$P_{fear} = \frac{E_{fear}}{E_{anger}+E_{anticipation}+E_{disgust}+E_{fear}+E_{joy}+E_{sadness}+E_{surprise}+}$$

Finally, the average of the percentage of each basic emotion was calculated, as shown in Table 2.

These values represent the emotional profile of each author in each platform.

## 4.3 Looking closer

The first step to analyse these results aims at determine the correlation between emotions from LinkedIn and Twitter, in order to identify differences between LinkedIn emotional profile and Twitter emotional profile. These results are presented in Table 3.

Considering that all LinkedIn messages are politically correct, when analysing the correlations between platforms, it evidences the proximity of emotional profile between LinkedIn and Tweeter of the authors involved with communication (C. Fairchild, J.Battelle and A. Mitchell) while for the other 3 authors the correlation is not so strong, allowing to distinguish the domain under which the author is writing.

In a second step, a new analysis was made concerned with emotional profile between authors in order to identify which authors are emotionally close to others according to the social media. It is expected that the proximity remains the same in different social media. For this objective, the emotional profile of all authors was correlated with each other for both social media. Table 4 shows the results of these correlations; the strongest are overlined and the weakest are underlined.

As we delve deeply into the analysis of correlations between authors, it is possible to highlight that in general, the correlations values are lower on Twitter when compared to LinkedIn, indicating a greater emotional distance between the authors on Twitter, reinforcing the

Table 2: Sentiment analysis from LinkedIn and Tweeter per author

| Author | Source | Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust |
|---|---|---|---|---|---|---|---|---|---|
| **C. Fairchild** | **LinkedIn** | 9% | 18% | 5% | 11% | 13% | 10% | 7% | 27% |
| | **Twitter** | 7% | 17% | 6% | 14% | 12% | 12% | 8% | 23% |
| **J. Saper** | **LinkedIn** | 7% | 20% | 3% | 10% | 14% | 8% | 8% | 32% |
| | **Twitter** | 5% | 22% | 4% | 7% | 20% | 5% | 13% | 24% |
| **J. Battelle** | **LinkedIn** | 12% | 17% | 8% | 14% | 11% | 10% | 5% | 23% |
| | **Twitter** | 11% | 17% | 6% | 12% | 13% | 9% | 10% | 22% |
| **B. McGowan** | **LinkedIn** | 11% | 18% | 6% | 15% | 10% | 11% | 8% | 21% |
| | **Twitter** | 9% | 18% | 6% | 11% | 16% | 10% | 9% | 20% |
| **A. Mitchell** | **LinkedIn** | 6% | 15% | 3% | 8% | 9% | 6% | 5% | 48% |
| | **Twitter** | 3% | 21% | 2% | 6% | 17% | 4% | 10% | 38% |
| **L. Profeta** | **LinkedIn** | 9% | 16% | 8% | 13% | 14% | 13% | 8% | 20% |
| | **Twitter** | 9% | 18% | 2% | 5% | 25% | 6% | 7% | 28% |

Table 3: Correlation between LinkedIn emotional profile and Twitter emotional profile

| Author | $r^2$ |
|---|---|
| C. Fairchild | 0.96 |
| J. Saper | 0.86 |
| J. Battelle | 0.92 |
| B. McGowan | 0.81 |
| A. Mitchell | 0.93 |
| L. Profeta | 0.79 |

idea of a common emotional profile in LinkedIn like a "common mask for everyone" and a "emotional freedom" in Tweeter.

### 4.4 Machine learning analysis

In order to achieve the objective of identifying the domain where the text was written, it was used an approach using machine learning for classification based on emotions contained in the text as model's dimensions. For this purpose, it was created a new dataset based on the preprocessed information from the emotional analysis. Each line of this dataset contains 11 dimensions, referring to the eight basic emotions according to the Plutchik [9] model, the polarities (positive and negative) and the source of information (social media name) regarding to the preprocessed text.

Later, this dataset was loaded and tested using several different algorithms in order to classify the social media according the emotions.

In our tests, the best classification score was obtained using a Random Forest algorithm, using a 10 fold cross-validation for training and testing, which achieved an weighted average of 96.21% of correct classified instances, which considers number of instances of each class for the weights, as presented in Table 5.

## 5 Conclusion

This paper presents a combination of lexicon-based and machine learning approaches to explore the emotions contained in a text through practices in sentiment analysis in order to detect the emotional profile and predict the conversation environment/domain.

Based on the analysis presented, it is possible to claim that there is an emotional profile according to each domain (in this case Twitter and LinkedIn). If this emotional profile is known, it is possible to adapt the discourse according to the audience, so that there is an emotional levelling of the author's discourse to their audience. This means that it is relevant to identify that context, or discourse environment, from the analysis of the communicator emotional writing style. So systems like chatbots can adapt their emotional profile according to the interactions received from people or even other systems, interacting with the user - at least emotionally - like a human.

As future work, it is planned to expand this analysis to include the emotional intensity profile, by combining with other text analysis metrics, in order to increase the emotional profile identification.

## Acknowledgements

This work has been supported by COMPETE: POCI-01-0145-FEDER-0070 43 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/ 00319/2013.

## REFERENCES

[1] Matteo Baldoni, Cristina Baroglio, Viviana Patti, and Paolo Rena. From tags to emotions: Ontology-

Table 4: Correlation between authors according social media

| | | C. Fairchild | J. Saper | J. Battelle | B. McGowan | A. Mitchell | L. Profeta |
|---|---|---|---|---|---|---|---|
| **LinkedIn** | **C. Fairchild** | 1.00 | **0.99** | 0.92 | **0.91** | 0.95 | 0.95 |
| | **J. Saper** | **0.99** | 1.00 | 0.90 | **0.89** | 0.94 | 0.93 |
| | **J. Battelle** | 0.92 | 0.90 | 1.00 | **0.95** | **0.86** | 0.92 |
| | **B. McGowan** | 0.91 | 0.89 | **0.95** | 1.00 | **0.81** | 0.92 |
| | **A. Mitchell** | **0.95** | 0.94 | 0.86 | 0.81 | 1.00 | **0.84** |
| | **L. Profeta** | **0.95** | 0.93 | 0.92 | 0.92 | **0.84** | 1.00 |

| | | C. Fairchild | J. Saper | J. Battelle | B. McGowan | A. Mitchell | L. Profeta |
|---|---|---|---|---|---|---|---|
| **Twitter** | **C. Fairchild** | 1.00 | 0.76 | **0.92** | 0.90 | 0.87 | **0.73** |
| | **J. Saper** | **0.76** | 1.00 | 0.85 | **0.94** | 0.92 | 0.92 |
| | **J. Battelle** | 0.92 | **0.85** | 1.00 | 0.94 | **0.95** | 0.86 |
| | **B. McGowan** | **0.90** | **0.94** | 0.94 | 1.00 | 0.93 | 0.94 |
| | **A. Mitchell** | **0.87** | 0.92 | **0.95** | 0.93 | 1.00 | 0.89 |
| | **L. Profeta** | **0.73** | 0.92 | 0.86 | **0.94** | 0.89 | 1.00 |

Table 5: Data classification results

| TP Rate[4] | FP Rate[5] | Precision | Recall | F-Measure | Class |
|---|---|---|---|---|---|
| 0.870 | 0.011 | 0.959 | 0.870 | 0.913 | LinkedIn |
| 0.989 | 0.130 | 0.963 | 0.989 | 0.976 | Twitter |
| **0.962** | **0.103** | **0.962** | **0.962** | **0.962** | **Weighted Avg.** |


driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, 6(1):41–54, 2012.

[2] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11:450–453, 2011.

[3] COBUILD Advanced English Dictionary. Politically correct definition | Collings English Dictionary.

[4] Paul Ekman. Basic emotions. *Handbook of cognition and emotion*, 98:45–60, 1999.

[5] Ingo Feinerer, Kurt Hornik, and David Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, March 2008.

[6] Matthew L. Jockers. *Syuzhet: Extract Sentiment and Plot Arcs from Text*, 2015.

[7] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[8] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. 29(3):436–465, 2013.

[9] Robert Plutchik. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219, 1984.

[10] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 516–527, 2016.

[11] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM, 2008.

[12] Gerhard Widmer. Tracking context changes through meta-learning. *Machine Learning*, 27(3):259–286, 1997.