Determining emotional profile based on microblogging analysis

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

Keywords: Sentiment Analysis, Emotion Analysis, Natural Language Processing, Social Media

Abstract: In general, groups of people are formed because of the similarities and affinities that members have with each other. Musical preferences, soccer teams or even similar behaviours are examples of similarities and affinities that motivate group formation. In social media, identifying these affinities is a difficult task because personal information is not easily identified. In this paper we present an alternative to identifying similarities between authors and their most frequent audience in Twitter, using emotional and grammatical writing style analysis. Through this study it is possible to define the creation of an emotional profile entirely based on the interactions of people, thus allowing software like chatbots to "learn emotions" and provide emotionally acceptable responses.

1 Introduction

Probably, one of the well known and used proverbs is: "Birds of a feather, flock together". However, what does it mean? In general meaning, it refers that people with common traits, interests and tastes tend to associate and relate with each other, in the same way as birds of the same species flock together. It can be observed in several different human behaviours, where people with common personalities tend to relate to each other.

Psychodynamic researchers claim that personality structure is set in childhood. For Sapir [24], the individual personality is formed around 2 or 3 years old, mostly through child training practices. Freud [3] argues that when the Oedipal complex is resolved, all basic structures of personality - the id, ego, and superego - are fully developed in opposition to Erikson [2] and Loevinger [10], which believe that personality continues to develop later in life. Sharing the same vision of Erikson and Loevinger, the motivational speaker Jim Rohn [23] claimed that "you are the average of the five people you spend the most time with".

Through social media usage - in general microblogging - people (authors) can express their opinion, desires and thoughts to a broad audience - from friends to unknown followers - keeping proximity despite physical distance. However, is this audience interested in the author's posts because they share the same sentiment, mood or emotions? Also, since software has no childhood, neither id, ego, and superego, is it possible to create an emotional profile based on existing ones, enabling the software "learn" how to have a personality?

In this paper, we present an approach for emotional profile creation based on existent emotional profiles, using emotion-based analysis to determine the proximity of the author's emotional and grammatical writing style with their audience on microblogging.

The remainder of this paper is as follows: Section 2, introduces the concept of emotion and presents some theories 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 used in our analysis and presents some data regarding them. Section **??** discusses about the data obtained and their impact, and finally, the paper ends in Section 5 with the conclusion and future work.

2 Emotion theories

In the literature, there are several models that attempt to explain the emergence of emotions and their associated behaviours. The main research theories here surveyed to serve as background to our analytical work are discrete, dimensional and appraisal theories.

Discrete emotional theories propose the existence of basic emotions that are universally displayed and recognized, grouped into categories and independent. An example of discrete emotional theory is proposed by Plutchik [21], where all sentiment is composed of a set

of 8 basic emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*).

On the other hand, **dimensional theories** characterize emotions regarding two or three dimensions, generally "arousal" and "valence." Valence is related to a positive or negative evaluation and is associated with the feeling state of pleasure (vs displeasure). Arousal reflects the general degree of intensity felt. However, using this two-dimensional is confusing to different emotions that share the same values of valence and arousal, as *anger* and *fear*. For this reason, it is common to add a third dimension to support this differentiation, as intensity. According to Leventhal [8] "the third view emphasises the distinct component of emotions, and is often termed the componential view."

Emotional-cognitive psychologists focus their studies mainly on the **appraisal process**. According to Scherer [25], the central idea is that emotions are triggered and differentiated by subjective analysis of an event, situation or object. For instance, Bill and Mike are watching a football game where their teams are playing. Bill's favourite team wins (event). Mike's appraisal is that an undesirable event happened. For Bill, the appraisal is that the event is desirable. So, the same event has produced opposite appraisals. In fact, emotions are triggered by the personal interpretation of the annoying or cheerful aspects of an event, the appraisal.

3 Related work

Due to the extensive usage, sentiment analysis on microblogs can be considered an opinion-rich resource and has been gaining popularity and attracting researchers from other areas to correlate information about specific events (e.g. Christmas, football matches, elections) with the sentiment contained in posts.

To perform sentiment analysis on microblogging, according to Pang et al.,[20], a straightforward approach is to exploit traditional sentiment analysis models. However, such methods are inefficient because they ignore some unique characteristics of microblog's data, as emoticons representations. Moreover, there are lots of colloquial terms, abbreviations and misspelt words used in microblogs which leads to heavy preprocessing tasks in order to identify its occurrences and "translate" them to a canonical form to be interpreted correctly. Due to such properties, several models have been developed especially for microblogs sentiment analysis recently.

An example of correlations between events and sentiments was proposed by Bollen et al. [1], which measures the sentiments on Twitter during a period and compares the correlation between sentiments contained in the text and significant events, including the stock market, elections and Thanksgiving. Also, Kim et al. [7] examined a dataset containing tweets about Michael Jackson's death in order to analyse how emotion is expressed on Twitter. O'Connor et al.[18] have analysed the sentiments about politicians, detecting a strong correlation between the aggregated sentiment and manually collected poll ratings.

Hu et al. [6] predict the individual well-being, as measured by a life satisfaction scale, through the language people used on social media. This is made using randomly selected posts from Facebook and a lexicon-based approach to identify the text words polarities.

A different approach of sentiment analysis using Twitter posts was presented by Pak and Paroubek [19], which consists of a linguistic analysis of the collected corpus to build a sentiment classifier. This classifier can determine positive, negative and neutral sentiments for a document.

Go et al. [5] proposed a framework which interprets the emoticons in tweets as noisy labels using supervised learning. However, as Liu [9] describes, there are some disadvantages when using only emoticons as noisy labels. A reason for this is because it is difficult to collect a large number of tweets with emoticons because they are time-related, dynamic and region-related. For Lu [12], "usually we can only exploit topic-independent tweets with emoticons. That is to say, in topic-dependent datasets which focus on one given topic, the performance boost brought by emoticons is not significant enough. Besides emoticons, rich topic-dependent unlabelled data can be exploited better."

Despite vast works about sentiment analysis in microblogs, none concerns on the study of the relationship between emotional profile and the writing style similarities among authors and their audiences.

4 Data analysis

In order to analyse the correlation between author's emotional writing style and their audience, we collected 2500 recent Twitter tweets from 6 different aleatory authors from different areas, as presented in Table 1:

Although it is clear that three authors do not post in Twitter - i.e. it is a press office representing them - the idea is of this paper is analyse the emotional, grammatical and textual proximity between authors and audiences, even if an author and/or an audience is a press office. In a different point of view, it can highlight an "press office emotional style", which can inform even where the conversation has occurred, as presented by Martins [15].

Table 1: Tweets authors

Author	Area	Origin
Elon Musk	Business	Press office
Katy Perry	Entertainment	Press office
Donald Trump	Business	By himself
Alan Shipnuck	News	By himself
Michele Dauber	Education	By herself
Floyd Mayweather	Sports	Press office

All tweets were gathered using the package TwitteR [4] for R [22]. Additionally, the tweets were labelled with an annotation indicating if the message was produced by a press office or by the author himself. During the gathering processes, we considered only the tweets and discarded the re-tweets. This decision wad adopted to avoid that texts from other author, like digital influencers or unknown viral texts, biased the individual analysis.

The task of analysing the emotional profile can be split into some intermediate steps: first, it was necessary for some preprocessing tasks in order to reduce data size by removing unnecessary text from the original message; Later, the relevant remaining text was analysed in order to evidence the author's polarities and the author's emotional style.

4.1 Preprocessing

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

In our analysis, the preprocessing pipeline begins with tokenization and in subsequent starts three parallel jobs, as shown in Fig. 1: Part of Speech Tagging (POS-T), Named Entity Recognition (NER) and Stopwords Removal. This strategy was used because both POS-T and NER need the text in the original format, in order to return the correct data from the analysis.

The POS-T process identifies the text grammatical structure and preserves nouns, verbs, adverbs and adjectives. The reason for this approach is because only these grammatical categories can bring emotional information. In a formal description, the Tokenization process converts the original text D in a set of tokens $T = \{t_1, t_2, ..., t_n\}$ where each element contained in T is part of the original document D. Later, the POS-T labels each token with semantic information and the process keeps all nouns, verbs, adverbs and adjectives in a set set P_T, where $P_T = \{p_{T_1}, p_{T_2}, ..., p_{T_k}\}$ and $0 \leq$ k $\leq$ n and $P_T \subset T$.

Like POS-T, NER process identifies names in 3 different categories: "Location", "Person" and "Organization" and removes all tokens related with these categories. As a result, a set $N_T = \{n_{(T_1)}, n_{(T_2)}, ..., n_{(T_j)}\}$ is constructed based on identified word category where $\forall j, cat(N_j) = "O"$. 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).

The Stopwords list is a predefined set $SW = \{sw_1, sw_2, ...sw_y\}$ of words, available in R through the package **tm**[16]. This step will return a set $T' = t'_1, t'_2, ..., t'_n$, where $T' \cap SW = \emptyset$.

After the 3 preprocessing tasks finish, the outcoming set ST is defined as $ST = T' \cap P_T \cap N_T$.

Later, a stemming algorithm is responsible for obtaining the stem of a word. For this task, we adopted an implementation of the Lovins stemmer [11], resulting in a set of stemmed words $PR = \{ST_1, ST_2, ..., ST_z\}$ ready to be analysed.

For all three tasks - POS-T, NER and Tokenization - the Stanford Core NLP [13] toolkit was used. An example of how the steps change the information is presented in Fig. 2.

4.2 Polarity analysis

In order to determine the author's polarity style, after the preprocessing all sentences contained in PR were compared against EmoLex lexicon [17] in order to identify the positive and negative sentiment of the entire text. Later, it was collected and analysed tweets from the top 5 most contacted audience from the author, in order to analyse the proximity of their polarities tweets and author, as presented in Table 2.

When applying the Pearson's correlation coefficient (r^2) between polarities authors and their respective top audience give $r^2 = 1$ for all results, indicating a **very strong** correlation between author's polarities and top audience's polarities.

Another analysis made was creating the sets:

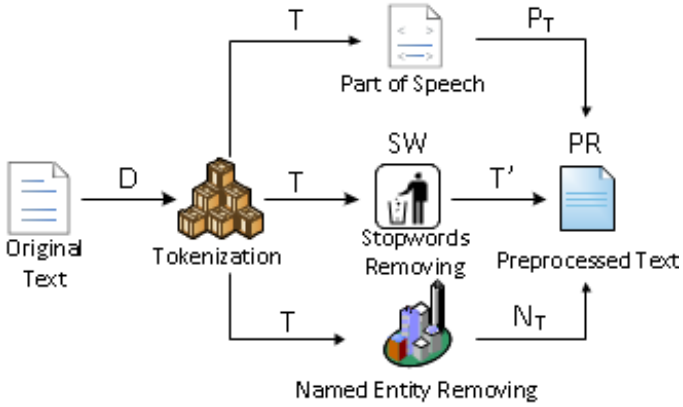

Figure 1: Preprocessing tasks

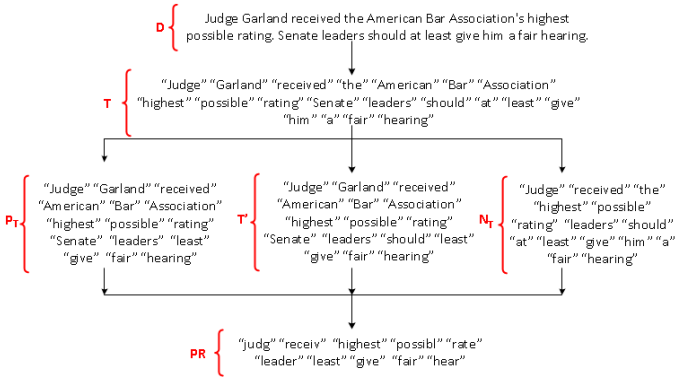

Figure 2: Preprocessing example

Table 2: Polarities analysis from authors and their top 5 audience

	Author		Audience's average	
	Positive	**Negative**	**Positive**	**Negative**
Elon Musk	0,68	0,26	0,68	0,28
Donald Trump	1,13	0,76	1,05	0,57
Katy Perry	1,01	0,27	0,59	0,18
Alan Shipnuck	0,45	0,27	0,57	0,38
Michele Dauber	0,68	0,65	0,83	0,70
Floyd Mayweather	0,62	0,13	0,61	0,27

$A = \{a1, ..., a6\}$ of authors,
$AP = \{ap_{a_1}, an_{a_1}, ..., ap_{a_6}, an_{a_6}\}$ of polarities where ap is the author positive polarity, an is the author negative polarity,
$C_A = \{c_{a_1,1}, ..., c_{a_i,j}\}$ of author's topmost contacts, where $0 \leq i \leq 6$ and $1 \leq j \leq 5$,
$CP = \{\overline{cp_i}, \overline{cn_i}\}$ of polarities, where $\overline{cp}$ is the average of audience's positive polarities, $\overline{cn}$ is the average of audience's negative polarities and $0 \leq i \leq 6$.

When applying the correlation coefficient between AP and CP, the result is $r^2 = 0,85$, indicating a strong correlation between authors' polarities their audiences polarities.

4.3 Emotional analysis

In order to analyse the emotions contained into the text, it was used a lexicon-based approach provided by Syuzhet package in R, in order to identify the emotions contained in text according to the model proposed by Plutchik [21], where all sentiment is composed of a set of 8 basic emotions (*anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise* and *trust*).

After all author's tweets analysis, the distribution of each basic emotion introduces a specific emotional profile for each author, defined as "emotional writing style", which is presented Table 3.

Using this information, the next step was to determine the average of each basic emotion for the audi-

ence's author. To achieve this objective, we used the same strategy used for the polarities, resulting in the audience's emotional writing style, according to Table 4.

Hence, when applying the Pearson's correlation coefficient (r^2) between basic emotions from authors and the average of their audiences, it is possible to verify that in significant part they are strongly correlated, as presented in Table 5.

Moreover, in order to establish Jim Rohn's statement, the analysis was expanded to verify the emotional profile of 100 most frequent contacts from each author. According to Fig. 3, the correlation between authors' emotions and most frequent contacts emotions' average decreases when the number of contacts increases, supporting that "you are the average of 5 people you spend the most time with."

A new point identified during the analysis, as showed in Table 6, is that the correlations tend to be higher within authors from the same origin, indicating a "press office emotional pattern."

4.4 Grammatical analysis

Another approach used was to determine if both authors and audiences share the same grammatical style when writing. For grammatical style, we understand the distribution of grammatical categories of words in sentences. To achieve this objective, both authors and audiences had their tweets labelled according to Part of Speech Penn Treebank [14] tags using Stanford Core NLP [13]. The next step was to determine the average of each Part of Speech tag for each author and their audience, resulting in the grammatical style, according to Table 7.

Hence, when applying the Pearson's correlation coefficient (r^2) between the grammatical style of authors and their audience, it is possible to verify that they are strongly correlated, as presented in Table 8.

4.5 Similarity analysis

The objective of the similarity analysis is to quantify the level of similarity of the author's texts and their respective audiences' texts. For this analysis, we collected the last 1000 tweets for each author and the last 1000 tweets for the same audiences used in section 4.2 in order to identify the similarity between their texts.

Before analysing the texts, they were preprocessed using the same pipeline described in Section 4.1 in order to keep the texts in the same structure in for the different analysis.

Using the Jaccard distance as metric to analyse the similarity among the texts; initially, we analysed the similarity between each author's texts and the texts of all audiences, in order to identify which audience is more similar to the author. Once identified the text's similarity percentage, we calculated the average of each audience, according to the formula:

$$\frac{\sum_{i=1}^{n} SM1_i + \sum_{i=1}^{n} SM2_i + \sum_{i=1}^{n} SM3_i + \sum_{i=1}^{n} SM4_i + \sum_{i=1}^{n} SM5_i}{n}$$

Later, for each author, we calculated the mean and standard deviation for the similarity between him and the audiences, as presented in Table 9.

This information allowed to identify that, in most cases, the highest similarity average was between the author and his audience. Moreover, the cases where it did not occur, the similarities values of the author's audience added with standard deviation indicates that the audience's value is close to the highest value.

5 Conclusion

This paper presented an analysis of the emotional and grammatical writing styles similarity from authors and their most frequent audiences on microblogs. This approach used lexicon-based techniques to explore the emotions contained in tweets and NLP techniques to identify grammatical excerpts.

Once the emotional and grammatical writing styles have very high values, indicating a strong correlation between authors and audiences, it is possible to conclude that both authors and audiences share the same writing style. Moreover, the correlation between authors emotions and the most frequent audiences emotions exhibited in Fig. 3 is high, and as the size of this audience increases, the lower the correlation becomes, confirming Jim Rohn's claiming.

This is a crucial issue because it enables the possibility of chatbots to create an emotional profile based on the interactions received from people or even other systems, creating an identity and interacting with the end user in smooth communication. Combining Generative Adversarial Networks (GANs) and emotional profiles, a new generation of chatbots can create its own "personality" and generate textual responses that fit its emotional profile.

As future work, it is planned to create a running example of an emotional chatbot that "learns" the emotional profile from the interactions and communicates according to this profile.

Table 3: Basic emotions per author

Author	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Elon Musk	0,11	0,34	0,06	0,14	0,23	0,10	0,12	0,38
Donald Trump	0,34	0,53	0,20	0,36	0,45	0,40	0,33	0,83
Katy Perry	0,12	0,52	0,04	0,13	0,67	0,16	0,25	0,43
Alan Shipnuck	0,11	0,20	0,09	0,14	0,20	0,15	0,11	0,26
Michele Dauber	0,37	0,32	0,23	0,34	0,20	0,31	0,10	0,48
Floyd Mayweather	0,11	0,44	0,02	0,10	0,43	0,08	0,18	0,37

Table 4: Basic emotion's frequency average of audience's author

Average audience of	Anger	Anticipation	Disgust	Fear	Joy	Sadness	Surprise	Trust
Elon Musk	0,13	0,38	0,08	0,20	0,20	0,13	0,13	0,30
Donald Trump	0,25	0,48	0,12	0,36	0,40	0,34	0,33	0,80
Katy Perry	0,10	0,35	0,05	0,09	0,42	0,11	0,18	0,34
Alan Shipnuck	0,18	0,31	0,13	0,22	0,25	0,21	0,18	0,33
Michele Dauber	0,45	0,36	0,26	0,53	0,22	0,36	0,17	0,64
Floyd Mayweather	0,19	0,34	0,08	0,14	0,35	0,12	0,16	0,33

Table 5: Correlation between basic emotion's authors and frequency average basic emotion's top audiences

Author	Correlation	Author	Correlation
Elon Musk	0,93	Donald Trump	0,99
Katy Perry	0,99	Alan Shipnuck	0,96
Michele Dauber	0,95	Floyd Mayweather	0,98

Table 6: Correlation between basic emotion's authors

	Elon Musk	Donald Trump	Katy Perry	Alan Shipnuck	Michele Dauber	Floyd Mayweather
Elon Musk	1,00	0,92	0,77	0,94	0,48	0,89
Donald Trump	0,92	1,00	0,61	0,95	0,64	0,71
Katy Perry	0,77	0,61	1,00	0,79	-0,04	0,97
Alan Shipnuck	0,94	0,95	0,79	1,00	0,49	0,84
Michele Dauber	0,48	0,64	-0,04	0,49	1,00	0,11
Floyd Mayweather	0,89	0,71	0,97	0,84	0,11	1,00

Table 7: Grammatical style for authors and audiences

	Part of Speech	Elon Musk	Donald Trump	Kate Perry	Alan Shipnuck	Michele Dauber	Floyd Mayweather
Author	CC	0,38	0,61	0,48	0,24	0,46	0,26
	CD	0,27	0,28	0,2	0,12	0,15	0,48
	DT	0,91	1,66	1,08	0,96	1,32	0,68
	IN	1,12	2,07	1,37	0,87	1,39	0,9
	JJ	1,05	1,39	1,09	0,83	1,11	0,63
	JJR	0,06	0,07	0,04	0,02	0,06	0,02
	JJS	0,06	0,07	0,03	0,03	0,01	0,04
	MD	0,28	0,31	0,1	0,15	0,21	0,05
	NN	2,94	3,18	3,58	2,56	3,73	2,72
	NNS	0,61	0,96	0,96	0,48	0,69	0,39
	POS	0,03	0,05	0,06	0,08	0,07	0,02
	RB	0,94	0,93	0,57	0,65	0,92	0,29
	RP	0,03	0,06	0,03	0,04	0,02	0,03
	TO	0,3	0,58	0,22	0,18	0,38	0,29
	VB	0,64	0,89	0,48	0,43	0,77	0,66
	VBD	0,18	0,4	0,25	0,25	0,46	0,12
	VBG	0,25	0,51	0,54	0,19	0,4	0,18
	VBN	0,21	0,35	0,12	0,12	0,26	0,05
	VBP	0,31	0,43	0,69	0,31	0,55	0,23
	VBZ	0,38	0,47	0,49	0,37	0,55	0,2
	WP	0,04	0,12	0,14	0,03	0,06	0,02
	WRB	0,03	0,05	0,06	0,04	0,11	0,03
Audience	CC	0,29	0,35	0,15	0,26	0,35	0,25
	CD	0,46	0,23	0,25	0,38	0,25	0,31
	DT	0,95	1,21	0,66	1	1,39	0,93
	IN	1,33	1,67	0,98	1,13	1,61	1,07
	JJ	1,02	0,9	0,62	0,84	1,13	0,82
	JJR	0,03	0,06	0,03	0,05	0,06	0,03
	JJS	0,05	0,05	0,03	0,03	0,03	0,03
	MD	0,2	0,14	0,05	0,18	0,2	0,13
	NN	3,65	3,93	2,92	2,96	3,68	3,33
	NNS	0,62	0,91	0,42	0,58	0,77	0,51
	POS	0,11	0,18	0,05	0,09	0,11	0,08
	RB	0,7	0,39	0,43	0,72	0,94	0,56
	RP	0,06	0,08	0,08	0,06	0,06	0,08
	TO	0,28	0,48	0,24	0,26	0,44	0,27
	VB	0,63	0,67	0,44	0,52	0,83	0,55
	VBD	0,27	0,32	0,14	0,37	0,35	0,25
	VBG	0,3	0,41	0,22	0,28	0,42	0,26
	VBN	0,23	0,28	0,09	0,19	0,25	0,17
	VBP	0,41	0,32	0,26	0,4	0,5	0,25
	VBZ	0,44	0,52	0,23	0,43	0,65	0,32
	WP	0,05	0,09	0,04	0,05	0,09	0,06
	WRB	0,08	0,06	0,05	0,08	0,11	0,05

Table 8: Correlation of grammatical frequency average between authors and audience

Author	Correlation	Author	Correlation
Elon Musk	0,99	Donald Trump	0,95
Katy Perry	0,98	Alan Shipnuck	0,99
Michele Dauber	0,99	Floyd Mayweather	0,99

Table 9: Similarities between authors and audiences

	Audiences							
Author	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**Mean**	**Standard Deviation**
Elon Musk (1)	**0,681%**	0,605%	0,629%	0,650%	0,542%	0,536%	0,607%	0,058%
Donald Trump (2)	0,728%	**1,068%**	0,840%	0,833%	0,893%	0,687%	0,842%	0,135%
Katy Perry (3)	0,712%	0,613%	0,999%	0,811%	0,623%	**1,096%**	0,809%	0,201%
Alan Shipnuck (4)	0,553%	0,531%	**0,738%**	0,664%	0,557%	0,573%	0,603%	0,081%
Michele Dauber (5)	0,739%	0,745%	0,731%	0,824%	**0,926%**	0,672%	0,773%	0,089%
Floyd Maywheather (6)	0,605%	0,542%	0,846%	0,566%	0,473%	**1,430%**	0,744%	0,360%

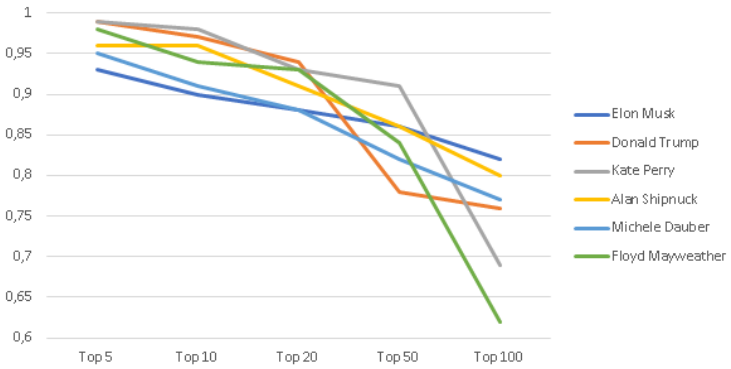

Figure 3: Correlations by author

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2019

REFERENCES

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

[2] Erik H Erikson. *Childhood and society*. WW Norton & Company, 1993.

[3] Sigmund Freud. *The ego and the id*. WW Norton & Company, 1962.

[4] Jeff Gentry. *twitteR: R Based Twitter Client*, 2015. R package version 1.1.9.

[5] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(2009):12, 2009.

[6] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.

[7] Elsa Kim, Sam Gilbert, Michael J Edwards, and Erhardt Graeff. Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter. *Web Ecology*, 3:1–15, 2009.

[8] Howard Leventhal and Klaus Scherer. The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and emotion*, 1(1):3–28, 1987.

[9] Kun-Lin Liu, Wu-Jun Li, and Minyi Guo. Emoticon smoothed language models for twitter sentiment analysis. In *Aaai*, 2012.

[10] Jane Loevinger. The meaning and measurement of ego development. *American Psychologist*, 21(3):195, 1966.

[11] Julie Beth Lovins. Development of a stemming algorithm. *Mech. Translat. & Comp. Linguistics*, 11(1-2):22–31, 1968.

[12] Tao-Jian Lu. Semi-supervised microblog sentiment analysis using social relation and text similarity. In *Big Data and Smart Computing (BigComp), 2015 International Conference on*, pages 194–201. IEEE, 2015.

[13] 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.

[14] Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[15] Ricardo Martins, José João Almeida, Pedro Henriques, and Paulo Novais. Domain identification through sentiment analysis. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 276–283. Springer, 2018.

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

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

[18] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11(122-129):1–2, 2010.

[19] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[20] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

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

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[23] E James Rohn and E James Rohn. *The treasury of quotes*. Health Communications, 1994.

[24] Edward Sapir. Personality//encyclopedia of the social sciences. *Seligman E., Johnson A*, pages 85–88, 1934.

[25] Klaus R Scherer, Tim Dalgleish, and Mick Power. Handbook of cognition and emotion. *Handbook of cognition and emotion*, 1999.