App2Check: a Machine Learning-based system for Sentiment Analysis of App Reviews in Italian Language

Emanuele Di Rosa, Alberto Durante
Head of Machine Learning & Semantic Analysis at Finsa spa, Assistant Research Scientist at Finsa spa
Via XX Settembre 14, Genova, Italy
E-mail: emanuele.dirosa@finsa.it, alberto.durante@finsa.it

Abstract

Sentiment Analysis has nowadays a crucial role in social media analysis and, more generally, in analysing user opinions about general topics or user reviews about product/services, enabling a huge number of applications. Many methods and software implementing different approaches exist and there is not a clear best approach for Sentiment classification/quantification. We believe that performance reached by machine learning approaches is a key advantage to apply to sentiment analysis in order to reach a performance which is very close to the one obtained by group of humans, who evaluate subjective sentences such as user reviews. In this paper, we present the App2Check system, developed mainly applying supervised learning techniques, and the results of our experimental evaluation, showing that App2Check outperforms state-of-the-art research tools on user reviews in Italian language related to the evaluation of apps published to app stores.

Keywords: Sentiment Analysis, Machine Learning, User Reviews, Italian Language, App2Check, iFeel, SentiStrength.

1. Introduction

Sentiment Analysis has nowadays a crucial role in social media analysis and, more generally, in analysing user opinions about general topics or user reviews about product/services, enabling a huge number of applications. For instance, sentiment analysis can be applied to monitoring the reputation or opinion of a company or a brand with the analysis of reviews of consumer products or services [1]. Moreover, it can also provide analytical perspectives for financial investors who want to discover and respond to market opinions [2,3]. Another important set of applications is in politics, where marketing campaigns are interested in tracking sentiments expressed by voters associated with candidates [4]. Sentiment analysis can also be applied to social platforms to show in real-time what is the opinion of people about emerging events and, in general, named entities, and about the relationships with other events and sources of information. In [5] it is also shown that the growth on the number of searches on the topic according to Google Trends, appears mainly after the popularization of online social networks.

App stores can be seen as another, not yet well explored, field of application of sentiment analysis. Indeed, they are another social media where users can freely express their own opinion through app reviews about a product, i.e. the specific app under evaluation, or a service, to which the considered app is connecting the user (e.g., a mobile banking app connects users to mobile banking services). In addition, reading user reviews on app stores shows that people frequently talk about and evaluate also the brand associated to the app under review: thus, it is possible to extract people opinion about a brand or the sentiment about a company or the provided service quality.

In this paper, we focus on the app store as a social media platform and on the sentiment evaluation in app reviews, which are examples of reviews related to a product, or a service or the associated brand. App reviews are a very interesting application in our opinion because they have not been extensively explored yet [6], and also because the sentiment score detected in a comment can significantly differ from the score assigned by the user to the app under evaluation. For example, a user can assign his good score to the app (i.e. assigning 5 stars) but also express in natural language some suggestions or highlight some –even important– bugs that, if they may not influence the user overall app evaluation, from the perspective of the developer or app producers are very important. For example, the comment[1] *"Great app to be honest, but it freezes while scanning the code of the pre-printed payment slip, it crashes, and it closes. Do something!!"* was rated 4 stars by the user. However, the comment describes a severe bug that causes an app crash and we can agree that this comment has an overall negative sentiment, especially from the perspective of developers. Vice versa, the user can assign, in general, a low rating but highlight some good features. All of this non-structured information is fully missing by only superficially evaluating an app through a 1 to 5 overall score –or any other product evaluated by the user with both sentences and a score –.

About the methods of processing user reviews, many methods and software implementing different approaches exist and there is not a clear best approach for Sentiment classification/quantification [7,8,9]. In [5] it is also shown that more than 7,000 articles have been written about sentiment analysis applying different approaches or slightly different algorithms and various startups are developing tools and strategies to extract sentiments from text. From our side, we believe that performance reached by machine learning approaches is a key advantage to apply to sentiment analysis in order to reach a performance which is very close to the one obtained by group of humans evaluating subjective sentences such as user reviews.

[1] The original comment in Italian is "*Ottima app, per carità, ma effettuando i pagamenti bollettini premarcati si blocca con la scannerizzazione del codice, va in crash e si chiude. Fate qualcosa!!*".

In this paper, we present the App2Check system, developed mainly applying supervised learning techniques and focused – in this first release – on Italian language, and present the results of our experimental evaluation showing that App2Check (version 1.0) outperforms state-of-the-art research software on user reviews in Italian language related to apps (which are, at the moment, our main target application). We considered research tools for our experimental evaluation, since the current state-of-the-art commercial tools recently included strict restrictions related to the possibility to run them for competitive analysis or benchmarking. In particular, since there are not so many research tools managing natively the Italian language, we applied the approach already shown in [7] where the iFeel research platform has been presented. iFeel performs the promising approach to translate sentences into English before running 19 state-of-the-art research tools. In order to make a fair comparison, we also included in our comparison a research tool that natively manages the Italian language: to the best of our knowledge, it is the only research tool with this feature that is available for download.

The structure of the paper is the following. After the current introduction about the main paper topics, in section 2, we report a description of the research tools we used to perform the comparison. In section 3, we briefly describe our system App2Check; in section 4, we present and discuss our experimental evaluation and, in section 5, we provide the paper conclusions.

2. State-of-the art Research Tools

In this section, we describe the research tools that will be mentioned in the following sections and included in our experimental evaluation: iFeel, a platform developed at Federal University of Minas Gerais and running 19 research tools, and SentiStrength version for Italian language.

2.1 iFeel

iFeel is a research web platform [10] allowing to run 19 state-of-the art research tools for sentiment analysis on the specified list of sentences. It allows to natively run tools supporting English and to first translate sentences from other languages into English and then run the underlying tools on the English translated sentences. It has been experimentally shown in [7] that well known language specific methods do not have a significant advantage over a simple machine translation approach.

The tools included in iFeel are the following (in alphabetical order): AFINN, Emolex, Emoticon DS, Emoticons, Happiness Index, NRC Hashtag, Opinion Finder, Opinion Lexicon, Panas-t, SANN, SASA, Senticnet, Sentiment140, SentiStrength, SentiWordNet, SO-CAL, Stanford Deep Learning, Umigon, Vader. We report in the following a few sentences from [13] describing each tool included in iFeel, in order to give an insight of the techniques implemented in these tools.

2.1.1. AFINN

It is a lexicon-based approach described in [14] and uses a Twitter based sentiment lexicon including Internet slangs and obscene words. AFINN can be considered as an expansion of ANEW, a dictionary created to provide emotional ratings for English words. ANEW dictionary rates words in terms of pleasure, arousal and dominance.

2.1.2. Emolex

It is a lexicon-based approach described in [17]. It uses a general sentiment lexicon supported by crowdsourcing. Each entry lists the association of a token with 8 basic sentiments: joy, sadness, anger, etc.

2.1.3. Emoticon DS (Distance Supervision)

It is a lexicon-based approach described in [18] and creates a scored lexicon based on a large dataset of tweets. It is based on how the frequency each lexicon occurs with positive or negative emotions.

2.1.4. Emoticons

It is a lexicon-based approach described in [15] where messages containing positive/negative emoticons are simply associated to a positive/negative sentiment, respectively. Messages without emoticons are not classified.

2.1.5. Happiness Index

It is a lexicon-based approach described in [19] and consists in a measure evaluating the psychological valence (happiness) distribution for words in the Affective Norms for English Words (ANEW). For each text, it is thus possible to compute the weighted average of the valence of the ANEW study words in a given text.

2.1.6. NRC Hashtag

It is a lexicon-based approach described in [20] and it builds a lexicon dictionary using a Distant Supervised Approach. It uses known hashtags (i.e. #joy, #happy, etc.) to 'classify' the tweet. Afterwards, it verifies frequency each specific n-gram occurs in an emotion and calculates its Strength of Association with that emotion.

2.1.7. Opinion Finder

It performs both a lexicon-based and a machine learning-based approach, as described in [21]. It performs subjectivity analysis through a framework that applies before lexical analysis and then a machine learning algorithm.

2.1.8. Opinion Lexicon

It is a lexicon-based approach described in [22] and it focuses on product reviews. It builds a lexicon to predict the polarity of product features that are summarized to provide an overall score to that product feature.

2.1.9. Panas-t

It is a lexicon-based approach described in [23] that detects mood fluctuations of users on Twitter. The method consists of an adapted version (PANAS) Positive Affect Negative Affect Scale of a well-known method in psychology with a large set of words, each of them associated with one from eleven moods such as surprise, fear, guilt, etc.

2.1.10. SANN

It performs both a lexicon-based and a machine learning-based approach, as described in [24]. It infers additional user ratings by performing sentiment analysis (SA) of user comments and integrating its output in a nearest neighbor (NN) model.

2.1.11. SASA

It is a machine learning-based approach, as described in [25], and it detects public sentiments on Twitter during the 2012 U.S. presidential election. It is based on the statistical model obtained from the Naive Bayes classifier on unigram features. It also explores emoticons and exclamations.

2.1.12. Senticnet

It is a lexicon-based approach described in [26]. It applies dimensionality reduction to infer the polarity of common sense concepts and hence provide a resource for mining opinions from text at a semantic level.

2.1.13. Sentiment140

It is a machine learning-based approach described in [27]. Sentiment140 is an ensemble of three classifiers (Naive Bayes, Maximum Entropy, and SVM) built with a huge amount of tweets containing emoticons collected by the authors.

2.1.14. SentiStrength

It performs both a lexicon-based and a machine learning-based approach, as described in [11]. It uses a lexicon dictionary annotated by humans and improved with the use of machine learning. We provide more details in section 2.2.

2.1.15. SentiWordNet

It performs both a lexicon-based and a machine learning-based approach, as described in [28]. It uses a lexical resource for opinion mining based on WordNet. The authors grouped adjectives, nouns, etc. in synonym sets (synsets) and associated three polarity scores (positive, negative and neutral) for each one.

2.1.16. SO-CAL

It is a lexicon-based approach described in [29]. It creates a new lexicon with unigrams (verbs, adverbs, nouns and adjectives) and multi-grams (phrasal verbs and intensifiers) hand ranked with scale +5 (strongly positive) to –5 (strongly negative). The authors also included part of speech processing, negation and intensifiers.

2.1.17. Stanford Deep Learning

It is a machine learning-based approach described in [30]. It applies a model called Recursive Neural Tensor Network (RNTN) that processes all sentences dealing with their structures and compute the interactions between them. The RNTN approach takes into account the order of words in a sentence, which is ignored in most of the methods.

2.1.18. Umigon

It is a lexicon-based approach described in [31] that disambiguates tweets using lexicon with heuristics to detect negations plus elongated words and hashtags evaluation.

2.1.19. VADER

It is a lexicon-based approach described in [32]. It is a human-validated sentiment analysis method developed for Twitter and social media contexts. VADER was created from a generalizable, valence-based, human-curated gold standard sentiment lexicon.

2.2 SentiStrength for Italian Language

SentiStrength was produced as part of the CyberEmotions project, supported by EU FP7. It estimates the strength of positive and negative sentiment in short texts, even for informal language. According to the authors, it has human-level accuracy for short social web texts in English, except political texts [11]. SentiStrength authors make available a version of the tool which natively manages Italian language. All tests have been carried out with both average emotion and strongest emotion options, but in this paper we only report the results obtained with the latter option turned on, due to better performance. Since the English version of SentiStrength is also included in iFeel, we ran on our own the Italian version and we will call it in the following SentiStrengthIta.

3. App2Check system description

App2Check is our system using an approach in which supervised learning methods are applied in order to build a predictive model for sentiment *quantification*. The training of the model is performed by considering a huge variety of language domains and different kinds of user reviews. App2Check provides, as answer to a sentence in Italian language, a quantification of the sentiment polarity scored from 1 to 5, according to the most recent trend shown in the last sentiment evaluation SemEval [12], where tracks considering quantification have been introduced. Thus, we consider the following quantification: as “positive”, sentences with score 4 (positive) or 5 (very positive); as “negative”, sentences with score 1 (very negative) or 2 (negative); as “neutral”, sentences with score 3. In order to compute the final answer, App2Check does not use just the prediction coming from the predictive model, but it applies also a set of algorithms which take into account some natural language processing techniques, allowing e.g. to also automatically perform topic/named entity extraction.
It is not possible to give more details about the engine due to non-disclosure restrictions.

App2Check is not only constituted by a web service providing access to the sentiment prediction of sentences, but it is also a full user-friendly web application allowing (more features in next release) in the current release 1.0 to:

a) Search for the app a user wants to monitor on the Apple App store, Google Play store or Microsoft Marketplace
b) Show the main topics discussed in user reviews which are both comment-specific, associated to a specific month or evaluated to overall the app life
c) Show the sentiment about the former extracted topics, including in the topics –if discussed in user comments–

also the company brand and the provided service level

d) Show a sentiment comparison on the app time horizon between apps owned by different app publishers (even market competitors).

A demo of the App2Check is available after sending a request by email to the first author of the paper.

4. Experimental Evaluation

In our experimental evaluation we considered user reviews of apps from Apple App store and Google Play store. More specifically, we focused on two different sets of comments. Test set A is made of 10 thousands comments from 10 different very popular apps (one thousand comments per app). These comments are associated only to an overall score for the app, called app rating in the app stores. Test set B is made of 1 thousand comments from the famous Candy Crush Saga app: in this case, we performed a manual quantification (in the 1-5 range) of the sentiment (from now on called *human sentiment classification* or ***HSC***).

We ran App2Check, iFeel and SentiStrengthIta on these user reviews, in order to evaluate:

- on test set A, their relative performance using the app rating as a reference indicator, i.e. as an approximation of the user sentiment so that we avoid to manually classify the sentiment for 10 thousand comments. Of course, considering a single comment, as already said, in general, the score/rating expressed by a user respect to an app can be substantially different respect to the sentiment expressed by a human. However, we experienced that the average score/rating of many (hundreds of) comments can be an approximation of the average sentiment expressed by a human on the same set. In Table 1 we show this phenomenon: human sentiment classification (performed by only one person trained with guidelines and examples) agrees with rating on 79.8% of cases with app rating.

- on test set B, the performance of the three systems is compared respect to the sentiment manually classified/quantified by a person on 1 thousand reviews of Candy Crush app (his classification is made publicly available). Thus, in this case we compare systems on a reference that is not approximated.

All of the user reviews together with a limited demo access to the prediction web service, are made available by contacting the authors, in order to make the experiments repeatable.

	Precision	Recall	F1
Negative	79.2	92.7	85.4
Neutral	14.6	21.9	17.5
Positive	96.6	78.6	86.7

Table 1: Comparing HSC vs App Rating on **Candy Crush Saga app**: accuracy is **79.8%**, other measures in table.

4.1 Systems comparison on Candy Crush Saga app Reviews

Tool	MF1	Acc	F1(-)	F1(x)	F1(+)
App2Check	**59.2**	**78.3**	**79.2**	14.0	**84.4**
Umigon	47.5	54.2	54.5	**16.7**	71.4
SentiWordNet	47.4	62.5	65.3	6.3	70.7
Sentiment140	47.1	63.6	72.2	3.6	65.6
SentiStrength	46.6	56.8	45.1	18.0	76.7
AFINN	46.5	56.7	51.7	14.0	73.7
Stanford DL	45.9	54.1	62.3	10.6	64.8
Op. Lexicon	45.7	51.6	53.8	15.5	67.7
NRC Hashtag	44.7	59.1	68.6	3.9	61.7
Emolex	39.8	44.6	44.6	13.4	61.4
SASA	37.9	44.7	38.9	13.0	61.8
Vader	37.8	41.1	32.4	16.7	64.1
Senticnet	37.2	53.1	35.9	7.8	67.8
SO-CAL	36.6	39.4	44.4	12.7	52.8
SentiStrengthIta	34.1	39.8	31.3	11.2	59.9
H. Index	29.8	40.1	16.7	11.7	61.1
Emoticon DS	23.8	50.8	2.1	1.7	67.5
Op. Finder	21.2	20.6	22.4	13.8	27.3
SANN	12.8	14.3	1.6	14.0	22.8
Panas-t	5.7	8.1	1.6	13.7	1.8
Emoticons	0.0	7.4	-	13.7	-

Table 2: Comparison of the tools respect to app rating.

In all of the following tables we show macro F1 (MF1), accuracy (Acc), F1 on the negative class (F1(-)), F1 on the neutral class (F1(x)), and F1 on the positive class (F1(+)). We highlight in bold the best value per column. In Table 2 we compare the tools on test set B (1 thousand user reviews from the popular Candy Crush app) with respect to the app rating. It shows that App2Check has the highest macro F1 (59.2%) and the highest accuracy (78.3%), calculated using app rating as a reference. The second and third accuracy is obtained by Sentiment140 and SentiWordNet, respectively. SentiStrengthIta produced a bad performance with respect to the English version of the same tool. In Table 3, we make a comparison with respect to the human sentiment classification. App2Check wins again here, showing the highest macro F1 (65.8%) and accuracy (81.8%); we see that it is even higher than the one calculated in Table 2 using app rating as a reference. This indicates that *App2Check is closer to the human sentiment classification (which is our goal) than to just the app rating*. In Table 3 we can also see that Sentiment140 and SentiWordNet have the second and third macro F1 and accuracy, respectively. Almost all of the tools show the same pattern and we obtain almost the same chart, thus by confirming that, if we consider hundreds of comments, using app rating becomes –overall and on average– an approximation of the user sentiment. The latter result enables us to use app rating in the following experiments as a reference approximating the sentiment expressed by one single person on the test set.

Tool	MF1	Acc	F1(-)	F1(x)	F1(+)
App2Check	**65.8**	**81.8**	**85.9**	25.2	**86.4**
Sentiment140	58.1	71.6	80.4	21.4	72.5
SentiWordNet	57.2	67.3	73.5	**26.6**	71.4
Stanford DL	53.7	60.5	70.0	21.2	69.9
NRC Hashtag	52.9	65.5	76.4	16.4	66.0
Umigon	50.9	56.4	54.0	20.5	78.1
SentiStrength	50.4	57.8	47.5	25.6	78.0
Op. Lexicon	49.8	54.2	53.1	23.7	72.6
AFINN	49.7	57.4	52.4	21.8	74.9
Vader	40.9	42.8	31.9	22.8	68.1
Senticnet	40.2	50.6	37.4	19.7	63.6
Emolex	40.0	43.3	43.1	16.1	60.9
SASA	39.3	44.3	40.8	15.8	61.3
SO-CAL	39.0	40.8	45.2	17.0	54.8
SentiStrengthIta	38.2	41.5	34.3	20.5	59.7
H. Index	32.3	39.8	16.9	18.6	61.4
Op. Finder	23.6	22.9	24.9	18.1	27.9
Emoticon DS	21.5	41.6	1.8	4.1	58.7
SANN	16.5	17.8	1.8	19.9	27.7
Panas-t	7.4	11.0	1.3	18.7	2.2
Emoticons	0.0	10.3	-	18.6	-

Table 3: Comparison of the tools respect to HSC.

4.2 Systems comparison on 10 thousand user reviews from 10 different apps

In Table 4 we show the results of the systems on 10 thousand reviews, selected considering 1 thousand reviews per each of the following popular apps: Angry Birds, Banco Posta, Facebook, Fruit Ninja, Gmail, Mobile Banking Unicredit, My Vodafone, PayPal, Twitter, Whatsapp.

Tool	MF1	Acc	F1(-)	F1(x)	F1(+)
App2Check	**73.3**	**85.7**	**82.7**	**45.6**	**91.7**
SentiWordNet	47.9	65.9	60.4	6.2	77.1
AFINN	47.5	60.3	49.2	16.6	76.7
SentiStrength	47.5	59.7	46.3	19.3	76.8
Stanford DL	45.6	54.0	56.5	13.5	66.8
Op. Lexicon	44.9	55.3	45.1	17.5	72.2
Sentiment140	44.1	58.7	57.4	6.7	68.2
Umigon	42.8	50.1	47.8	14.6	66.2
SO-CAL	41.8	49.3	45.8	13.8	65.6
NRC Hashtag	41.2	52.9	53.6	8.3	61.7
Senticnet	40.9	63.1	36.6	9.1	76.9
Vader	38.5	46.2	29.5	19.7	66.3
Emolex	38.3	45.5	38.5	14.1	62.3
SASA	37.7	48.8	29.6	16.4	67.1
SentiStrengthIta	34.0	39.6	32.0	13.9	56.2
H. Index	31.1	39.0	21.9	13.4	57.9
Emoticon DS	27.8	63.6	3.0	2.5	77.8
Op. Finder	26.0	26.6	25.3	15.4	37.2
SANN	12.2	14.9	3.6	16.5	16.4
Panas-t	6.0	9.0	1.2	16.0	0.9
Emoticon	5.4	8.8	0.1	15.9	0.4

Table 4: Comparison of the tools on 10 thousand user reviews from 10 different apps respect to app rating.

Considering app rating as a reference, we clearly see that App2Check outperforms all of the other tools, reaching an accuracy of about 86%. In order to better analyze App2Check performance, in Figure 1 we show a plot of the average sentiment per month of all user reviews (1 for positive, 0 for neutral and -1 for negative sentiment). In the plot we include app rating as a reference, App2Check and SentiStrengthIta (SS. ITA in the plot), since they natively support Italian, and the two best tools according to accuracy from Table 4: SentiWordNet (SWN), which is also the best according to macro-F1, and Emoticon DS (Emo DS). Emoticon DS assign too often a positive score, in fact its graph is very close to 1, even where the average rating is negative. In fact, the high accuracy reached by this tool is a consequence of the number of positive documents in this testset. It is clear that the other tools follow quite well the trend of the rating plot. Both SentiStrengthIta and SentiWordNet, instead, are closer each other and to the app rating, but their evaluation is under the reference plot. App2Check is the closest to the app rating, but in certain areas it differs from the rating, especially when the score provided by the user is on average far away from the sentiment expressed in the review. In our opinion, App2Check would have even higher accuracy on these 10 thousand instances, considering as a reference the human sentiment classification: this is made clear while using the web application and evaluating the answer of the system on every single user comment.

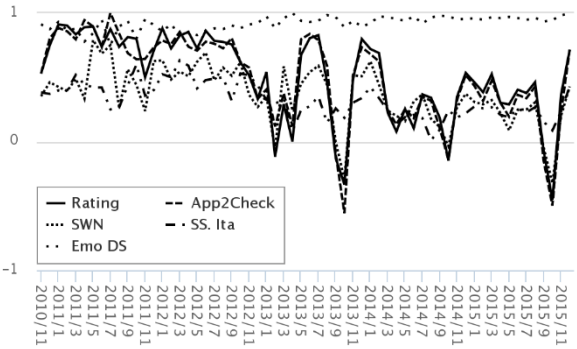

Figure 1: Comparison of 4 tools on 10 thousand chronologically sorted user reviews respect to app rating.

5. Conclusion and future work

In this paper we presented App2Check, a machine learning-based system performing sentiment classification or quantification on user reviews in Italian language. We evaluated it on 11 thousand user reviews related to apps published in app stores. Results show that App2Check outperforms state-of-the-art research tools on this test set. As future work, we want to extend the system to work on more languages and we want to extend the system evaluation on different kind of user reviews and on user feedbacks from Twitter.

6. Bibliographical References

[1] Hu, M., Liu, B. (2004): Mining and summarizing customer reviews. *In Proc. of KDD 2004*, pp. 168–177.

[2] Oliveira, N., Cortez, P., Areal, N. (2013). On the

predictability of stock market behaviour using stocktwits sentiment and posting volume. *In Proc. of LNCS 2013*, vol. 8154, pp. 355–365.

[3] Bollen, J., Mao, H., Zeng, X.-J. (2010). Twitter Mood Predicts the Stock Market. *CoRR* abs/1010.3003

[4] Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *In Proc. of ICWSM 2010*.

[5] Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., Benevenuto, F. (2015). A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods.

[6] Guzman, E., Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *In Proc. of RE 2014*, pp. 153–162.

[7] Araújo, M., Reis, J. C. S., Pereira, A. C. M., Benevenuto, F. (2016). An Evaluation of Machine Translation for Multilingual Sentence-level Sentiment Analysis. *In Proc. of ACM SAC 2016*, pp. 1140-1145.

[8] Esuli, A., Sebastiani, F., (2010). Al and Opinion Mining, Part 2. IEEE Computer Society, pp. 153–162.

[9] Gao, W., Sebastiani, F. (2015). Tweet Sentiment: From Classification to Quantification. *In Proc. of IEEE/ACM ASONAM 2015*, pp. 97–104.

[10] Araújo, M., Gonçalves, P., Cha, M., Benevenuto, F. (2014). ifeel: A system that compares and combines sentiment analysis methods. *In Proc. of ACM WWW 2014*, pp. 75–78.

[11] Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., Kappas, A. (2010). Sentiment strength detection in short informal text. *ASIST journal*, 61(12), pp. 2544–2558.

[12] SemEval-2016 Task 4: Sentiment Analysis in Twitter.

[13] Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science journal*, 5(1), pp. 1–29.

[14]. Nielsen F (2011) A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv:1103.2903

[15] Gonçalves P, Araujo M, Benevenuto F, Cha M (2013) Comparing and combining sentiment analysis methods. *In Proc. of COSN'13*.

[16] Hu M, Liu B (2004) Mining and summarizing customer reviews. *In Proc. of KDD 2004*, pp 168-177.

[17] Mohammad S, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. *In Computational Intelligence* 29(3):436-465.

[18] Hannak A, Anderson E, Barrett LF, Lehmann S, Mislove A, Riedewald M (2012) Tweetin' in the rain: exploring societal-scale effects of weather on mood. *In Proc. of ICWSM 2012*.

[19] Dodds PS, Danforth CM (2009) Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies* 11(4):441-456. doi:10.1007/s10902-009-9150-9.

[20] Mohammad S (2012) #emotional tweets. *In Proc. of SemEval 2012.*

[21] Wilson T, Hoffmann P, Somasundaran S, Kessler J, Wiebe J, Choi Y, Cardie C, Riloff E, Patwardhan S (2005) OpinionFinder: a system for subjectivity analysis. *In Proc. of HLT-Demo 2005*.

[22] Hu M, Liu B (2004) Mining and summarizing customer reviews. In: KDD'04, pp 168-177.

[23] Gonçalves P, Benevenuto F, Cha M (2013) PANAS-t: a psychometric scale for measuring sentiments on Twitter. arXiv:1308.1857v1

[24] N. Pappas and A. Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. *In Proc. of SIGIR 2013*.

[25] Wang H, Can D, Kazemzadeh A, Bar F, Narayanan S (2012) A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In Prof. of ACL system demonstrations 2012, pp 115-120.

[26] Cambria E, Olsher D, Rajagopal D (2014) SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. *In Proc. of AAAI 2014*, pp 1515-1521.

[27] Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using Distant Supervision. *In Processing 2009*, pp. 1-9.

[28] Baccianella S, Esuli A, Sebastiani F (2010) SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *In Proc. of LREC 2010*, pp 2200-2204.

[29] Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. In *Computational Linguistics* 37(2):267-307.

[30] Socher R, Perelygin A, Wu J, Chuang J, Manning CD, Ng AY, Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *In Proc. of EMNLP 2013*, pp 1631-1642

[31] Levallois C (2013) Umigon: sentiment analysis for tweets based on terms lists and heuristics. *In Proc. of SemEval 2013*, pp 414-417.

[32] Hutto C, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. *In Proc. of ICWSM 2014*.