Automatic Evaluation of Employee Satisfaction

Marco Piersanti **Giulia Brandetti** **Pierluigi Failla**

Data Modeling and Analysis – Enel Italia S.R.L.
Rome, Italy
{name}.{surname}@enel.com

Abstract

English. Human Resources are one of the most important assets in modern organizations. Their capability of facing employees' needs is critical in order to have an effective and efficient company, where people are the center of all business processes. This work is focused on developing new techniques that, leveraging a data driven approach, can help Human Resources to find a more precise employee satisfaction categorization, to easily identify possible issues and to act in a proactive fashion.

Italiano. *Le Risorse Umane sono una delle funzioni piú importanti nelle aziende moderne. La loro capacità di affrontare le necessità dei dipendenti è fondamentale per avere un'azienda efficiente, dove le persone sono al centro di tutti i processi di business. Il presente lavoro è focalizzato sullo sviluppo di nuove tecniche che, facendo leva su un approccio data driven, possano aiutare le Risorse Umane a dare una categorizzazione della soddisfazione dei dipendenti piú precisa, ad identificare piú facilmente possibili problemi condivisi e ad agire in maniera proattiva.*

1 Introduction

Every modern organization has a dedicated function which takes care of its employees, commonly called Human Resources (HR). HR duties are related to the capability of creating value through people, ensuring that everyone can express his own potential and has a productive and comfortable office environment.

Nowadays, HR can rely on data to create a new paradigm based on a *data driven* approach, where analysts can leverage data in order to get more complete, detailed and data-supported decisions.

Being able to monitor employees' engagement and satisfaction is critical in order to maintain a positive and constructive office environment. The benefit for the company is in the capability of retaining the best employees and keeping the overall workforce strong and motivated. Furthermore, recent surveys (Globoforce, 2015) show the issues that companies are facing when they try to do retention or improve engagement.

This paper is organized as follows. Section 2 presents a literature review on both themes of HR Management and text mining, Section 3 summarizes the motivations that drove the present study, Sections 4 and 5 discuss data and methodology, respectively, and Section 6 presents the results. Finally, Section 7 discusses the implications of the findings and further possible developments.

2 Related Works

Despite the great interest that is arising around the application of Data Science methods and Natural Language Processing (NLP) to HR problems, very few studies exist on the topic.

The entire field of corporate HR Management has been revolutionized by the pioneering work done by People Operations at Google (well described in Bock (2015)), that first put a spotlight on the benefits of having a more scientific and rigorous approach to these areas which have been traditionally more reluctant to adopt change.

Employee satisfaction has been linked to long-run stock returns (Edmans, 2011), consistently with human relations theories which argue that employee satisfaction brings a stronger corporate performance through improved recruitment, retention, and motivation. Furthermore, Moniz and Jong (2014) followed an interesting approach to link employee satisfaction and firm earnings, based on sentiment analysis of employees' reviews from the career community website www.glassdoor.com.

Text clustering, and more generally text classification, is a well established topic in the NLP research area (Sebastiani, 2002; Aggarwal and Zhai, 2012; Kadhim et al., 2014). The automated categorization of texts, although dating back to the early '60s (Maron, 1961; Borko and Bernick, 1963), went through a booming interest in the last twenty years, due to the explosion of the amount of documents available in digital form and the impelling need to organize them. Nowadays text classification is used in many applications, ranging from automatic document indexing and automated metadata generation, to document filtering (e.g., spam filters (Drucker et al., 1999)), word sense disambiguation (Navigli, 2009), population of hierarchical catalogs of Web resources (Dumais and Chen, 2000), and in general any application requiring document understanding.

Flourished in the last decade, sentiment analysis aims to classify the polarity of a given text – whether the expressed opinion in a document or a sentence is positive, negative, or neutral (Pang et al., 2002; Pang and Lee, 2008; Baccianella et al., 2010; Liu, 2012). The growing interest on the subject reflects on the success of the tasks of sentiment analysis on Twitter data at SemEval since 2013 (Rosenthal et al., 2014; Rosenthal et al., 2015; Nakov, 2016). Even if the driving language for most of those techniques is English, we started to see an increasing trend also in Italy (Basile and Nissim, 2013; Basile et al., 2014; Basile et al., 2015), confirming the great interest of the Italian NLP community in sentiment analysis techniques.

3 Task Description

Enel HR Business Partners' (HR-BPs) job consists in monitoring employees' well-being, acting when necessary to solve issues. In doing so, they periodically interview employees and register information about their satisfaction, motivation, work-life balance and other personal issues in textual notes.

Currently, employees are manually classified by HR-BPs in three main categories: *Demotivated*, *Neutral* and *Motivated*. Unfortunately, employee motivation is not a very reliable indicator of employee well-being, since it may mask an underlying dissatisfaction, or more generally the presence of issues that HR department should act on. Indeed, one can face several problems in the office everyday life but still be motivated. We therefore chose to consider the sentiment, as it shows through interviews, as a proxy of employee satisfaction.

With the present study, we aim to categorize employee satisfaction in a more detailed and automatic way, identifying common trends among employees and clustering them into groups that share similar problems. The goal is to help HR-BPs in having an overall view of their resources' mood and make effective adjustments in critical situations. It will also help in such situations when new HR-BPs take over a group of already interviewed resources, allowing them to have a clearer understanding of the employees and their criticalities without having to read all interviews.

For all the aforementioned reasons, we performed a classification of the interviews based on their sentiment (Section 5.1) prior to send them into the text clustering algorithm (Section 5.2). In the present study, we chose to focus only on negative moods, since they include the biggest issues HR should monitor. Nevertheless, the practical usage of this system involves the whole set of sentiment classes, since HR is interested in monitoring the entire workforce well-being and in following its evolution over time.

In choosing methods, we had to tackle the challenge to balance the scientific rigor and the need of ease of interpretation and communication to all actors involved in the process. We therefore chose to use well understood and controllable techniques, like *sentiment analysis* and *k-means clustering*.

4 Experiments and Data

4.1 Data Description

HR System Integration provided interviews data, a file containing 53k textual notes in more than 5 languages taken by HR-BPs during interviews. Interviews spanned approximately 1 year, from June 2015 to July 2016, and they were performed by 142 different HR-BPs.

For the present study, we focused only on Italian interviews (25k interviews) and selected a single interview for each employee (23k interviews), since in the few cases of repeated interviews texts were not relevant (e.g., "See previous interview").

Notes shorter than 5 words (the 5th percentile of the distribution of the number of words in each note) were considered irrelevant. As a result, in the present study we considered a dataset of 22k interviews.

4.2 Data Preprocessing

Data preparation includes removing punctuation, numbers and stop words (we removed 300 common Italian stop words, including some peculiar words that are not relevant in this context, like “Enel”, “colloquio”, etc.), changing letters to lower case and lemmatization (Schmid, 1994). We assumed all unrecognized words to be typos, and we corrected them by using a dictionary composed by 110k Italian words and 650 English words commonly used in business daily-life[1]. In order to have an effective correction, we used Optimal String Alignment distance (Brill and Moore, 2000) (OSA distance), an extension of Levenshtein distance that, together with insertion, deletion and substitution, includes transpositions among its allowable operations.

5 Model Description

5.1 Sentiment Analysis

We performed sentiment classification of texts by customizing and improving a publicly available lexicon[2]. In total, we used 3428 Italian labeled unigrams and 10451 bigrams, categorized as positive (4736), neutral (4367) or negative (4776) based on their polarity.

The sentiment classification model proposed in this paper is based on a score φ_{sent} that weights differently unigrams and bigrams with a factor α:

$$\varphi_{sent} = (1-\alpha)\cdot\varphi_{uni} + \alpha\cdot\varphi_{bi}$$

where $0 \leq \alpha \leq 1$, φ_{uni} is the difference between the number of positive and negative unigrams, normalized by the number of words in the text and φ_{bi} is the difference between the number of positive and negative bigrams, normalized by the number of bigrams in the text. Final sentiment was then calculated according to the formula

$$\text{Sent} = \begin{cases} +1 & \text{if } \varphi_{sent} > \theta \\ -1 & \text{if } \varphi_{sent} < -\theta \\ 0 & \text{otherwise.} \end{cases}$$

Model calibration (i.e. the choice of parameters α and θ) was performed by comparing model results with the ones produced by manually annotating a subset of 200 (randomly chosen) texts (*training set*): two judges classified texts independently and a third one solved the cases where there wasn’t agreement. Agreement between the two independent judges was measured by calculating Cohen’s Kappa ($\kappa = 0.6$).

We chose $\alpha = 0.7$ and $\theta = 0.0004$ so that accuracy, recall and precision of the sentiment model were maximized. Although we may have chosen to optimize parameters in order to maximize negative texts recognition, we chose to consider the overall accuracy on the three classes, because from a business perspective it is more valuable to monitor the entire workforce satisfaction and to follow its evolution over time. While for α we tried manually different settings, weighting more bigrams than unigrams, for θ we used the ROC curve and the area under it, picking the one with maximal sum of true-positive and false-negative values.

5.2 Text Clustering

For notes’ clustering, we focused only on those classified as negative from the sentiment model (Section 5.1).

Since we didn’t have a target variable to model (unsupervised classification), we chose to adopt the k-means clustering algorithm, using k-means++ technique to seed the initial cluster centers (Arthur and Vassilvitskii, 2007).

The clustering model was applied on the TF-IDF matrix, built with bigrams appearing in at least 2 documents. In this way, we reduced our dimensionality from the initial 37k bigrams to 5k. To calculate proximity among documents, we used cosine similarity.

Additionally, *Silhouette distance* has been chosen to select the best number of clusters: different models were computed by varying the number of clusters between 2 and 30 and the respective Silhouette scores were compared, fixing the number of clusters at 12 (corresponding to the highest score).

6 Results

The application of this sentiment model (Section 5.1) classified interviews in 3655 negatives, 956 neutrals and 17297 positives. As we can see in Table 1, sentiment classification is more clearly related to employee satisfaction than motivation classes provided by HR-BPs, although they some-

[1] https://github.com/napolux/paroleitaliane

[2] https://github.com/opener-project/public-sentiment-lexicons

Text (after preprocessing)	HR-BP Motivation	Sentiment
risorsa brillante neodirigente clima positivo ansioso molto positivo (*brilliant resource new executive positive mood anxious very positive*)	Motivated	+1
assenteista risorsa molto critico non riuscire nulla (*absentee very critical resource don't succeed in anything*)	Demotivated	-1
non valorizzare poco riconoscimento non potere rimanere (*don't valorize inadequate recognition can't stay*)	Motivated	-1
molto scontento non credere azienda reale meritocrazia interessare piano esodo (*very unhappy don't believe company real meritocracy interest retirement plan*)	Motivated	-1
stabile routinario non proattivo scarso impegno (*stable routine not proactive scarce effort*)	Neutral	-1
assumere direttamente assistente seguire particolare sicurezza vedere capo (*hire directly assistant follow particular safety see boss*)	Neutral	0

Table 1: Examples of sentiment classification and comparison with HR-BPs motivation classes.

True/Predicted	-1	0	1	All
-1	12	11	3	26
0	3	20	18	41
1	1	37	95	133
All	16	68	116	200

Table 2: Confusion matrix. True values here represent manually labeled texts.

times are aligned.

A different subset of 200 manually labeled texts (*test set*), labeled with the same methodology as described in Section 5.1, was used for evaluating model performance. Accuracy and recall were both 64%, while precision was 70%. For more details about the sentiment classification performance, see confusion matrix in Table 2.

The clustering algorithm was applied only on the 2392 negative interviews and it identified 8 clusters that we were able to precisely label, while for the remaining 4 clusters labeling was unfeasible (see Table 3). Labels were applied by manually looking at the most frequent bigrams within clusters, trying to identify common significant topics.

The most frequent identified issues preventing employee satisfaction were *health problems*, the will to *change activity*, *compensation* and the high *workload*. The most frequent bigrams for clusters 0–3 were not specific enough to lead to a precise labeling, since they refer to work activity and job in general and they don't focus on clear issues.

In Figure 1, we represented clustering results by means of t-SNE, a popular method for exploring high-dimensional data (Maaten and Hinton, 2008). By this mean, we reduced the high-dimensionality space of bigrams to an artificial two-dimensional space (since dimensions here don't have a real meaning, we excluded them from the plot). For the sake of clarity, we chose not to show unlabeled clusters; the resulting plot shows that clusters are well separated and on average quite dense.

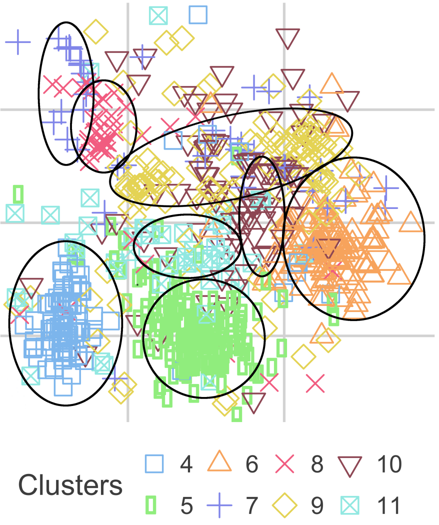

Figure 1: Clustering results represented with t-SNE. Only labeled clusters are shown.

7 Conclusions

The proposed approach could be a powerful tool for HR-BPs to better understand the main issues related to the lack of employees' satisfaction. Furthermore, it could help HR analysts to quickly decide which are the best actions to solve those issues, analyzing whether a complaint is isolated or shared by a group, whether it's trivial or urgent and act accordingly. As an example, HR Departments could test different actions over a group of unsatis-

Cluster id	Docs #	Label	Most frequent bigrams
0	382	(NA)	lavoro svolgere (*do work*)
1	76	(NA)	persona supporto (*support person*) supporto dipendente (*employee support*) carico lavoro (*workload*)
2	1985	(NA)	lavoro piacere (*enjoy work*)
3	33	(NA)	attività poco (*activity low*) solo attività (*only activity*) attività dovere (*activity must*)
4	149	Workload	carico lavoro (*workload*) eccessivo carico (*exaggerated load*) lamentare eccessivo (*complain about exaggerated*)
5	297	Health issues	problema salute (*health issue*) grave problema (*difficult problem*) serio problema (*serious problem*)
6	206	Change activity	cambiare attività (*change activity*) volere cambiare (*want to change*)
7	81	Low productivity	poco produttivo (*low productivity*)
8	67	Not productive	rispetto compito (*compliance with task*) compito non produttivo (*not productive task*)
9	173	Compensation	mancato riconoscimento (*lacking recognition*) lamentare mancato (*complain about lacking*)
10	134	Don’t change activity	svolgere attività (*do activity*) volere continuare (*want to go on*) continuare svolgere (*keep doing*)
11	72	Change job	cambio attività (*activity change*) cambiare lavoro (*change job*)

Table 3: Clustering results. Cluster id, number of documents within clusters, cluster labels and most frequent bigrams inside clusters are shown. Labels were applied by manually looking at the most frequent bigrams within clusters.

fied employees, in order to understand which one is the most effective for a given issue.

The very same model could also be used on neutral and positive subjects, so that HR could check whether the quality of life at work of these employees could be somehow improved, and understand which are the essential key factors for the employees’ well-being.

From a technical point of view, one possible improvement in order to strengthen the solidity of the present approach could be to manually annotate a subset of (anonymized) texts, developing a gold standard of HR interview clusters, to be used as a test set for techniques like the one presented in this study. This gold standard may be made available company-wise, in order to encourage collaboration and to foster the creation of a data science community, to help bring a data driven way of thinking even to those areas which have been traditionally more reluctant to adopt a rigorous digital transformation.

This is a first step to improve how HR Departments operate nowadays. We strongly believe that the introduction of a data driven approach can support critical HR decisional processes and improve companies’ productivity, without having to sacrifice each individual’s quality of life.

Acknowledgements

This research was supported by Enel. We thank our colleagues from HR System Integration dep. who provided the data analyzed in this study.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on italian tweets. In *Proceedings of the 4th*

Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 100–107.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the evalita 2014 sentiment polarity classification task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*.

Pierpaolo Basile, Valerio Basile, Malvina Nissim, and Nicole Novielli. 2015. Deep tweets: from entity linking to sentiment analysis. In *Proceedings of the Italian Computational Linguistics Conference (CLiC-it 2015)*.

Laszlo Bock. 2015. *Work rules!: Insights from inside Google that will transform how you live and lead*. Hachette UK.

Harold Borko and Myrna Bernick. 1963. Automatic document classification. *Journal of the ACM (JACM)*, 10(2):151–162.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 286–293. Association for Computational Linguistics.

Harris Drucker, Donghui Wu, and Vladimir N. Vapnik. 1999. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, 10(5):1048–1054.

Susan Dumais and Hao Chen. 2000. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 256–263. ACM.

Alex Edmans. 2011. Does the stock market fully value intangibles? employee satisfaction and equity prices. *Journal of Financial Economics*, 101(3):621–640.

Globoforce. 2015. 2015 employee recognition report – culture as a competitive differentiator. Technical report.

Ammar Ismael Kadhim, Yu-N Cheah, and Nurul Hashimah Ahamed. 2014. Text document preprocessing and dimension reduction techniques for text document clustering. In *2014 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, pages 69–73.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Melvin Earl Maron. 1961. Automatic indexing: an experimental inquiry. *Journal of the ACM (JACM)*, 8(3):404–417.

Andy Moniz and Franciska Jong. 2014. Sentiment analysis and the impact of employee satisfaction on firm earnings. In *Proceedings of the 36th European Conference on IR Research on Advances in Information Retrieval - Volume 8416*, ECIR 2014, pages 519–527, New York, NY, USA. Springer-Verlag New York, Inc.

Preslav Nakov. 2016. Sentiment analysis in twitter: A semeval perspective. In *Proceedings of NAACL-HLT*, pages 171–172.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 73–80. Dublin, Ireland.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 154–164.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.