Atlantis Studies in Uncertainty Modelling, volume 1

11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Sentiment Analysis in Social Networks: A Methodology Based on the Latent Dirichlet Allocation Approach

Fabio Clarizia, Francesco Colace, **Francesco Pascale**, **Marco Lombardi**, **Domenico Santaniello**
DIIn, Università degli Studi di Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano SA
email: {fclarizia, fcolace, fpascale, malombardi, dsantaniello}@unisa.it

Abstract

The detection and analysis of sentiment in textual communication is a topic attracting attention in both academia and industry. In fact, thanks to the explosion of the Social Networks a wealth of information is produced every day. This huge amount of contents can be very helpful in assessing the general public's sentiment and opinions toward products, services and topics. This paper presents a methodology for the detection of sentiment in textual contents using a methodology based on the Latent Dirichlet Allocation (LDA) approach and a word-based graphical model, the mixed graph of terms. The method has been tested in various operative scenarios: on standard datasets, on datasets obtained collecting tweets from twitter and on datasets coming from social networks as Twitter and TripAdvisor. The experimental campaigns show that the proposed approach is effective and furnishes good and reliable results in each context.

Keywords: Sentiment analysis, Knowledge Management, Information Retrieval, LDA Approach.

1 Introduction

Between the birth of the world and 2003, there were just a few dozen exabytes of information on the Web. Today, users create this amount of information weekly. Therefore, with the increasing availability of user-generated content, thanks to consumer opinion websites, social networks, blogs, forums, users have opportunities to express opinions and make them available to everyone. In this sense, a new collective intelligence is arisen and can be used as source of valuable information for the decision-making processes [1], [2], [3]. This is nothing new: the preparatory process for a choice has always been characterized by the search for information that could support it. This is even more true when you want to buy a product or service, especially online. In general, textual information can be divided in two main categories: facts and opinions. Facts are objective statements while opinions reflect people's sentiments about products, other person and events [4], [5], [6]. The interest, that potential customers show in online opinions and reviews about products, is something that vendors are gradually paying more and more attention. Companies are interested in what customers say about their products and how they can improve their appeal. Therefore, there is a lot of information on the web that have to be properly managed in order to provide vendors with highly valuable network intelligence and social intelligence to facilitate the improvement of their business.

In this scenario, a real interest is growing in the Sentiment Analysis research field. Sentiment Analysis, also known as Opinion Mining, is an attempt to take advantage of the vast amounts of user generated contents for the identification of the agreement or disagreement statements that deal with positive or negative feelings in comments or reviews [7],[8]. In literature there are several examples of sentiment mining approaches from textual content, which will be described in detail.

In this paper, we focus on a methodology for mining sentiment from texts by the adoption of a Latent Dirichlet Allocation Based Approach. Starting from the results obtained by the use of the LDA it is possible the extraction of a graph, named Mixed Graph of Terms, from a set of documents belonging to a same knowledge domain [9], [10], [11].

This graph contains a set of weighted word pairs, which can be discriminative for sentiment classification [12] because the LDA-based topic modeling is essentially an effective conceptual clustering process that helps the discovering of semantic features as affective relationships.

The organization of this paper is the following: in section 2, s

ome related works will be discussed and in the section 3, will introduce the Mixed Graphs of Terms, its main features and its building process. In section 4, the sentiment extraction approach will be discussed while in section 5 the experimental results will be described. In particular, the experimental campaign focuses on three different domains: a standard dataset, a dataset containing Tweets and a dataset obtained from a consumer opinion website (TripAdvisor).

2 Related works

With the advent of Web 2.0, to comment in real-time experiences with products or in certain contexts has become one of the services most preferred by users. The major companies, especially those active on the web, provide to its customers to post comments for evaluating the experience with the purchased products. The comment left on-line typically contains the feeling of the person towards products and services and may be a useful assessment tool for companies. Those on-line reviews, if properly analyzed, can provide vendors highly valuable network and social intelligence for improving their business. Many studies have focused on the economic values of reviews, exploring the relationships between the sales, the performance of products and their reviews [13], [14], [15], [16], [17], [18], [19]. In general, what the people thinks about a product can influence how well it sells, because these reviews represent the “wisdom of crowds” that is an effective indicator of the product’s future sales performance [20]. These considerations have inspired research in opinion mining for automatically detecting emotions, opinions and general evaluations from texts. In general, sentiment detection techniques can be divided in two main streams: lexicon based methods [21] and machine learning methods [22]. The lexicon based methods rely on a sentiment lexicon (e.g. a collection of known and pre-compiled sentiment terms). In literature there are many examples of this approach [23]. In [24] there is a detailed description of methods and techniques which adopt a sentiment lexicon for mining the sentiment inside a document. In general, there are three options for acquiring a sentiment lexicon: manual approaches, dictionary-based approaches and corpus-based approaches. According to a manual approach people code the lexicon by hand, while the dictionary-based approaches allow expanding a set of seed words by utilizing resources like Wordnet. In a corpus-based approach a set of seed words is expanded by using a large corpus of documents from a single domain. The manual approach is not feasible as each domain requires its own lexicon and such a laborious effort is prohibitive, so there are few examples in literature. More interesting is the dictionary-based approach where the starting point is a set of seed sentiment words which is expanded using a Wordnet’s synonyms and antonyms. Various algorithms have been developed and interesting approaches are in [25],[26]. The main problem with this kind of approach is that the acquired lexicon could not capture the specific peculiarities of a specific domain. In this case an effective solution to this problem is the construction of a domain specific sentiment lexicon by the use of a corpus-based algorithm. A first examples of this approach is in [27] where an algorithm for clustering, according to a set of linguistic connectors, adjectives that have a consistent polarity. More interesting and active is the research field related to the Machine Learning approaches that make use of syntactic or linguistic features, also if hybrid approach are very common [28], with a sentiment lexicon that still continue to play a key role [29]. In [30] three machine learning approaches (Naıve Bayes, Maximum Entropy and Support Vector Machines) has been adopted to label the polarity of a movie reviews datasets. A promising approach is presented in [31] where a novel methodology has been obtained by the combination of rule based classification, supervised learning and machine learning. In [32] a SVM based technique has been introduced for classifying the sentiment in a collection of documents. In [33], instead, a Naive Bayes classifier is used for the sentiment classification of tweets’ corpora. One of the most interesting approaches to the Sentiment Analysis is the statistical approach. These methods, by feeding a machine learning algorithm a large training corpus of affectively annotated texts, can learn not only the affective valence of affect keywords, but also to take into account the valence of other arbitrary keywords, punctuation and word co-occurrence frequencies. In general these methods are generally semantically weak and work with acceptable accuracy when given a sufficiently large text input [34]. Various papers face the sentiment analysis in this way: in [35] a probabilistic approach to sentiment mining is adopted: the Sentiment Probabilistic Latent Semantic Analysis (S-PLSA) in which a review, and more in general a document, can be considered as generated under the influence of a number of hidden sentiment factors. The S-PLSA is an extension of the Probabilistic Latent Semantic Analysis PLSA where it is assumed that there are a set of hidden semantic factors or aspects in the documents related to documents and words under a probabilistic framework. Several approaches adopt the LDA methodoly for text analysis [36],[37],[38],[39]. In general, these approaches use variations of LDA to uncover latent topics in a document collection: in this way the expectation is that these topics will correspond to rateable aspects for the document under review [40]. In [41] a hybrid HDP-LDA model aims to automatically determine the number of aspects, distinguish factual words from opinionated words and extracts the aspect specific sentiment words. In [42] a probabilistic modeling framework, called Joing Sentiment Topic (JST), based on LDA which detects sentiment and topic simultaneously from text has been introduced. In particular, this model extends the LDA sentiment mining approach by constructing an additional sentiment layer, assuming that topics are generated dependent on sentiment distributions and words are generated conditioned on the sentiment-topic pairs. In this paper, we investigate the adoption of an approach based on the Latent Dirichlet Allocation (LDA). As previously said, in LDA, each document may be viewed as composed by a mixture of various topics. This is similar to probabilistic latent semantic analysis, except that in LDA the topic distribution is assumed to have a Dirichlet prior. By the use of the LDA approach on a set of documents belonging to a same knowledge domain, a Mixed Graph of Terms can be automatically extracted. This graph contains a set of weighted word pairs that can be considered discriminative for sentiment classification. The main reason of such discriminative power is that LDA-based topic modeling is essentially an effective conceptual clustering process and it helps discover semantically rich concepts describing the respective affective relationships. By means of applying these semantically rich concepts, that contain more useful relationship indicators to identify the sentiment from messages and which allows to identify the kind of semantic relationship between word pairs in mGT, the proposed system can discover more latent relationships and make less errors in its predictions.

3 Mixed Graph of Terms

In this section we explain how a Mixed Graph of Terms can be extracted from a corpus of documents. The Feature Extraction module (FE) is represented in Fig. 1. The input of the system is a set of documents.

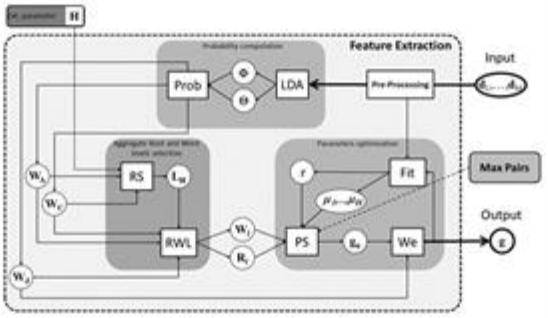

Figure 1: mGT Feature Extraction Process

After the pre-processing phase, which involves tokenization, stop words filtering and stemming, a Term-Document Matrix is built to feed the Latent Dirichlet Allocation (LDA) module. The LDA algorithm, assuming that each document is a mixture of a small number of latent topics and each word's creation is related to one of the document's topics, provides as output two matrices, which express probabilistic relations between topic-document and word-topic respectively. Under particular assumptions LDA module's results can be used for determining: the probability for each word occurring in the corpus (W_A); the conditional probability between word pairs (W_C); the joint probability between word pairs (W_J). Details on LDA and probability computation can be found on [43]. Defining Aggregate roots (A_R) as the words whose occurrence is most implied by the occurrence of other words of the corpus, a set of H aggregate root $r=(r_1,\ldots,r_H)$ can be determined in the following way:

$$r_i = \text{argmax}_{v_i} \prod_{j \neq i} P(v_i|v_j)$$

This phase is referred as Root Selection (RS) in Fig. 1. A weight ψ_{ij} can be defined as a degree of probabilistic correlation between AR pairs: $\psi_{ij} = P(r_i, r_j)$. We define an aggregate as word vs having a high probabilistic dependency with an aggregate root ri. Such a dependency can be expressed through the probabilistic weight $\rho_{is} = P(r_i|v_s)$. Therefore, for each aggregate root, a set of aggregates can be selected according to the highest weight values. As result of the Root-Word Level selection (RWL), an initial mGT structure, composed by H aggregate roots R_l linked to all possible aggregates W_l, is obtained. An optimization phase allows neglecting weakly related pairs according to fitness function. In particular, the proposed algorithm, given the number of aggregate roots H and the desired max number of pairs as constraints, chooses the best parameter settings τ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H)$ defined as follows:

- τ: the threshold that establishes the number of aggregate root/aggregate root pairs. A relationship between the aggregate root r_i and aggregate root r_j is relevant if $\psi_{ij} \geq \tau$
- μ_i: the threshold that establishes, for each aggregate root v_i, the number of aggregate root/word pairs. A relationship between the word v_s and the aggregate root r_i is relevant if $\rho_{is} \geq \mu_i$

A mixed graph of terms is then built from several clusters, each containing a set of words (aggregates) related to an (aggregate root), the centroid of the cluster. Some aggregate roots are also linked together building a centroids subgraph.

4 Searching the Sentiment by the use of the Mixed Graph of Terms

As described in the previous section a Mixed Graph of Terms represents a set of documents related to a well-defined knowledge domain by the identification of their most co-occurrent words. Thanks to the LDA approach, the Mixed Graph of Terms contains the words that better represent the knowledge domain and can be considered as a sort of filter for the classification of documents. In this paper we show how the Mixed Graph of Terms can be effectively applied for the sentiment mining from texts.

The architecture of the proposed approach is composed by two main modules:

- Mixed Graph of Terms building module: this module builds a Mixed Graph of Terms starting from a set of documents belonging to a well-defined knowledge domain and previously labeled according the sentiment expressed in them. In this way the obtained

Mixed Graph of Terms contains words, and their relative co-occurrences, that are representative of a certain sentiment in a well-defined knowledge domain. The output of this module is a sentiment oriented mixed graph of terms representing the documents and their sentiment. Thanks to the LDA approach the graph can be obtained by the use of a set of few documents [43]. In figure 1 the module architecture and its main functional steps are depicted.

- Sentiment Mining Module: this module extracts the expressed sentiment from a generic document. In particular, it compares the document to a sentiment oriented Mixed Graph of Terms belonging to the same knowledge domain. The input of this module are a generic document, a positive and negative sentiment oriented mixed graph of terms and the output is the sentiment detected in the input document. The sentiment extraction is obtained according to the following algorithm, which is reported in appendix A.

5 Experimental Results

The evaluation of the performances of the proposed approach is achieved through various experimental campaigns. For each of these experimental campaigns a specific dataset has been adopted aiming to point out the main characteristics of the proposed method.

In particular three operative scenarios have been set:

1. Movie Review Standard Dataset provided by Pang et al. [44]. This experimental campaign aimed to evaluate the approach's performance on a standard dataset. In particular, by this kind of experimentation a comparison with other methods that are in literature can be performed.
2. Twitter Dataset: in this experimentation the ability of the methodology to mine the sentiment in very short messages, as tweets, has been tested by the use of some datasets that have been built collecting tweets from well defined "hashtags". This kind of experimentation has been adopted also for evaluating the system's capability in the real time sentiment classification of tweets.
3. TripAdvisor Dataset: in this case some posts from the TripAdvisor platform has been collected and organized in a dataset. The aim of this experimentation was the evaluation of the system's performance depending on the length of the texts. In fact, in TripAdvisor the length of the various comments is highly variable and so an evaluation about the dependence between length of the comments and performance of the algorithm is really interesting.

The first dataset used for the experimentation is the Movie Reviews Dataset provided by Pang et al [44]. This dataset consist of 1000 positive and 1000 negative reviews from the Internet Movie Database. Positive labels were assigned to reviews that had a rating above 3.5 stars and negative labels were assigned to the rest. The first step of the experimental campaign aimed finding the best size for the training set. For achieving this task nine training sets have been built selecting in a random way from the 10% to 90% of the positive and negative comments that are in the full dataset. By the use of these training sets, the positive and negative mixed Graphs of Terms have been built and the sentiment classification on the remaining comments has been conducted. The second dataset used was built using the social network Twitter. Tweets are brief and public messages, which can be easily collected through an API service. Therefore, we collected 5000 tweets in Italian language related to cultural heritage filed hashtags. Each tweet has been labeled by experts according to its mood. Finally, another dataset was obtained using TripAdvisor. Comments of users released on this platform do not have a character limit, and express opinions about specific services. Through the API it was possible to collect 10000 users comments in English language related to the same topic of Twitter dataset.

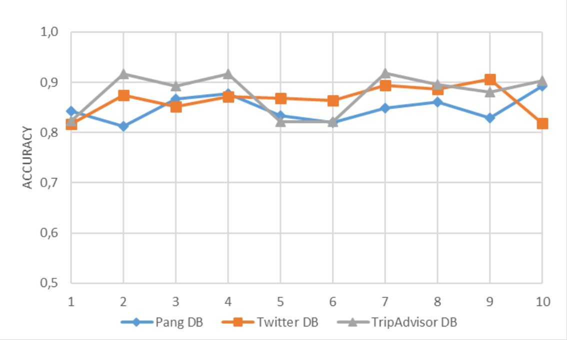

Figure 2: Accuracy Trend

Dataset	Accuracy Average Value
Pang	87,75%
Twitter	86,49%
TripAdvisor	**87,92%**

Table 1: Accuracy Average

For each of these datasets the process of training sets and mixed graph of terms building and documents classification has been conducted ten times. The obtained results, in terms of accuracy, are shown in figure 2. In Table 1 are reported the accuracy average of the proposed approach.

The proposed methodology presents satisfactory experimental results and shows the following characteristics:

- The proposed approach shows constant accuracy values regardless of the length of the comments and the number of words in it. This is not surprising as the proposed methodology is able to select the most significant words by linking the others to them. The strength of mGT is precisely in the possibility of representing a text through its main components.
- The method does not show significant variations in results depending on the language used both in terms of language and content. Once again the mGT is able to synthesize the document in its main components.
- The methodology shows a constant behavior in every situation. The experimentation, conducted with ten different experimental cycles, has always shown consistent and constant results.

The results obtained are in line with those present in the literature. Another specific characteristic lies in the possibility of continuously training the system through the analysis of new texts. This procedure allows a continuous updating with the relative possibility of adaptation to linguistic changes such as the introduction of new linguistic terms or patterns.

6 Conclusions

This paper introduces a novel methodology for sentiment mining by the using of the LDA approach. The sentiment of a text could be analysed thanks to a graphical structure named the Mixed Graph of Terms (mGTs) obtained analysing documents with the LDA approach. In this graphical structure there is a sort of digest of the documents main features. The proposed approach has been applied in three different scenarios. Three different dataset has been built collecting texts from the main social networks: Twitter and TripAdvisor. The prosposed approach shown indipendence from languages, topics and length of comments. The obtained results are promising and comparable with literature results. Future works aims to apply the proposed approach in real time scenarios.

References

[1] L. Garcia-Moya, H. Anaya-Sanchez, and R. Berlanga-Llavori, “Retrieving Product Features and Opinions from Customer Reviews,” *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 19–27, May 2013.

[2] F. Colace, L. Greco, S. Lemma, M. Lombardi, D. Yung, and S. K. Chang, “An Adaptive Contextual Recommender System: a Slow Intelligence Perspective,” 2015, pp. 64–71.

[3] F. Colace, M. Lombardi, F. Pascale, and D. Santaniello, “A multi-level approach for forecasting critical events in smart cities,” in *Proceedings - DMSVIVA 2018: 24th International DMS Conference on Visualization and Visual Languages*, 2018.

[4] B. Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions*. 2015.

[5] F. Colace, D. Santaniello, M. Casillo, and F. Clarizia, “BeCAMS: A behaviour context aware monitoring system,” in *2017 IEEE International Workshop on Measurement and Networking, M and N 2017 - Proceedings*, 2017.

[6] F. Clarizia, F. Colace, M. De Santo, M. Lombardi, F. Pascale, D. Santaniello, and A. Tuker, “A multilevel graph approach for rainfall forecasting: A preliminary study case on London area,” *Concurr. Comput. Pract. Exp.*, p. e5289, May 2019.

[7] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends® Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.

[8] F. Clarizia, F. Colace, M. Lombardi, F. Pascale, and D. Santaniello, *Chatbot: An education support system for student*, vol. 11161 LNCS. 2018.

[9] F. Colace, M. De Santo, L. Greco, and P. Napoletano, “Improving relevance feedback-based query expansion by the use of a weighted word pairs approach,” *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, pp. 2223–2234, Nov. 2015.

[10] M. Casillo, F. Clarizia, F. Colace, M. De Santo, M. Lombardi, and F. Pascale, “A Latent Dirichlet Allocation Approach using Mixed Graph of Terms for Sentiment Analysis,” in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[11] F. Clarizia, F. Colace, M. De Santo, M. Lombardi, and F. Pascale, “A Sentiment Analysis Approach for Evaluation of Events in Field of Cultural Heritage,” in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2018, pp. 120–127.

[12] F. Colace, L. Casaburi, M. De Santo, and L. Greco, “Sentiment detection in social networks and in collaborative learning environments,” *Comput. Human Behav.*, vol. 51, pp. 1061–1067, Oct. 2015.

[13] Y. Liu, X. Yu, X. Huang, and A. An, “Blog Data Mining: The Predictive Power of Sentiments,” in *Data Mining for Business Applications*, Boston, MA: Springer US, pp. 183–195.

[14] G. D’Aniello, M. Gaeta, and M. Z. Reformat, “Collective Perception in Smart Tourism Destinations with Rough Sets,” in *2017 3rd IEEE International Conference on Cybernetics (CYBCONF)*, 2017, pp. 1–6.

[15] E. Kalampokis, E. Tambouris, and K. Tarabanis, “Understanding the predictive power of social media,” *Internet Res.*, vol. 23, no. 5, pp. 544–559, Oct. 2013.

[16] G. Guzmán, “Internet search behavior as an economic forecasting tool: The case of inflation expectations,” *Journal of Economic and Social Measurement*. 2011.

[17] Y. Liu, Y. Chen, R. F. Lusch, H. Chen, D. Zimbra, and S. Zeng, “User-generated content on social media: Predicting market success with online word-of-mouth,” in *IEEE Intelligent Systems*, 2010.

[18] S. Brody, “An Unsupervised Aspect-Sentiment Model for Online Reviews,” *acl*, 2010.

[19] D. Maynard, D. Dupplaw, and J. Hare, “Multimodal sentiment analysis of social media,” in *CEUR Workshop Proceedings*, 2013.

[20] X. Yu, Y. Liu, X. Huang, and A. An, “Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 4, pp. 720–734, Apr. 2012.

[21] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, “Lexicon-Based Methods for Sentiment Analysis,” *Comput. Linguist.*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[22] E. Boiy and M.-F. Moens, “A machine learning approach to sentiment analysis in multilingual Web texts,” *Inf. Retr. Boston.*, vol. 12, no. 5, pp. 526–558, Oct. 2009.

[23] A. Montoyo, P. Martínez-Barco, and A. Balahur, “Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments,” *Decis. Support Syst.*, vol. 53, no. 4, pp. 675–679, Nov. 2012.

[24] R. Feldman, “Techniques and applications for sentiment analysis,” *Commun. ACM*, vol. 56, no. 4, p. 82, Apr. 2013.

[25] M. M. R. J. M. M. D. R. Jaap Kamps, “Using wordnet to measure semantic orientation of adjectives,” *ACL ’10 Proc. 48th Annu. Meet. Assoc. Comput. Linguist.*, 2010.

[26] E. C. Dragut, C. Yu, P. Sistla, and W. Meng, “Construction of a sentimental word dictionary,” in *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM ’10*, 2010, p. 1761.

[27] V. Hatzivassiloglou and K. R. McKeown, “Predicting the semantic orientation of adjectives,” in *Proceedings of the 35th annual meeting on Association for Computational Linguistics* -, 1997, pp. 174–181.

[28] G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall, “Predicting Emotional Responses to Long Informal Text,” *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 106–115, Jan. 2013.

[29] L. Qi and L. Chen, “Comparison of Model-Based Learning Methods for Feature-Level Opinion Mining,” in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2011, pp. 265–273.

[30] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?,” in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - EMNLP ’02*, 2002, vol. 10, pp. 79–86.

[31] R. Prabowo and M. Thelwall, “Sentiment analysis: A combined approach,” *J. Informetr.*, vol. 3, no. 2, pp. 143–157, Apr. 2009.

[32] K. P. P. Shein, “Ontology Based Combined Approach for Sentiment Classification,” in *Proceedings of the 3rd International Conference on Communications and Information Technology*, 2009, pp. 112–115.

[33] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, “Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1555–1565.

[34] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, “Statistical Approaches to Concept-Level Sentiment Analysis,” *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 6–9, May 2013.

[35] T. Hofmann, “Probabilistic Latent Semantic Analysis,” in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[36] Blei. David M., A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, 2003.

[37] W. Lu, J. Yang, and X. Liu, “The PID controller based on the artificial neural network and the differential evolution algorithm,” *J. Comput.*, vol. 7, no. 10 SPL.ISS., pp. 2368–2375, 2012.

[38] C. Lin, Y. He, R. Everson, and S. Ruger, “Weakly Supervised Joint Sentiment-Topic Detection from Text,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[39] I. Titov and R. McDonald, “A joint model of text and aspect ratings for sentiment summarization,” in *ACL: HLT*, 2008.

[40] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, “Multi-aspect Sentiment Analysis with Topic Models,” in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 81–88.

[41] W. Ding, X. Song, L. Guo, Z. Xiong, and X. Hu, “A Novel Hybrid HDP-LDA Model for Sentiment Analysis,” in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, 2013, pp. 329–336.

[42] C. Lin, Y. He, R. Everson, and S. Ruger, “Weakly Supervised Joint Sentiment-Topic Detection from Text,” *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[43] F. Colace, M. De Santo, L. Greco, and P. Napoletano, “Text classification using a few labeled examples,” *Comput. Human Behav.*, vol. 30, pp. 689–697, Jan. 2014.

[44] B. Pang and L. Lee, “Opinion Mining and Sentiment Analysis,” *Found. Trends® Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.

Appendix A: Sentiment_Mining_Algorithm

Input:

- $W = [w_1, w_2, \ldots, w_N]$ the N words that are in a Document D belonging to a knowledge domain K;
- The sentiment oriented mixed graphs of terms mGT_+ and mGT_- obtained analyzing documents related to the knowledge domain K;
- $RW_+ = [rw_{1+}, rw_{2+}, \ldots, rw_{t+}]$ the aggregator words that are in mGT_+;
- $PRWi_+ = [prw_{i1+}, prw_{i2+}, \ldots, prw_{i(t-1)+}]$ the values of co-occurences among the aggregator word rw_{i+} and the other aggregator words that are in the mGT_+
- $AW_{i+} = [aw_{i1+}, aw_{i2+}, \ldots, aw_{im+}]$ the cluster i of m aggregated words related to the rw_{i+} aggregator word that is in mGT_+;
- $PAW_{i+} = [paw_{i1+}, paw_{i2+}, \ldots, paw_{im+}]$ the values of co-occurences among the aggregated words and the rw_{i+} aggregator word in the mGT_+
- $RW_- = [rw_{1-}, rw_{2-}, \ldots, rw_{(c-1)-}]$ the aggregator words that are in mGT_+;
- $PRWi_- = [prw_{i1-}, prw_{i2-}, \ldots, prw_{i(c-1)-}]$ the values of co-occurences among the aggregator word rw_{i-} and the other aggregator words that are in the mGT_-
- $AW_{i-} = [aw_{1-}, aw_{2-}, \ldots, aw_{h-}]$ the cluster i of h aggregated words related to the rw_{i-} aggregator word that is in mGT_-;
- $PAW_{i-} = [paw_{i1-}, paw_{i2-}, \ldots, paw_{ih-}]$ the values of co-occurences among the aggregated words and the rw_{i-} aggregator word in the mGT_-
- L an annotated lexicon.

Output: $Sentiment_D \in$ {Positive, Negative, Neutral} the sentiment expressed in the document D

Algorithm Description

$f_p=0$ // index of positive sentiment of the document
$f_n=0$ // index of negative sentiment of the document
$f_{pre+}, f_{agg+}, f_{wag+}=0$
$f_{pre-}, f_{agg-}, f_{wag-}=0$

Determining the synonyms for each word belonging to the vector W

$RW_+ = RW_+ + Synset[L, RW_+]$ //The function Synset adds to RW_+ the synonymous of each word rw_{i+}, retrieved in the annotated lexicon L. For each synonymous of the word rw_{i+} a new $PRW_{h+} = [prw_{i1+}, prw_{i2+}, \ldots, prw_{i(h-1)+}]$ is considered.

$RW_- = RW_- + Synset[L, RW_-]$ //The function Synset adds to RW_- the synonymous of each word rw_{i-}, retrieved in the annotated lexicon L. For each synonymous of the word rw_{i-} a new $PRW_{m-} = [prw_{i1-}, prw_{i2-}, \ldots, prw_{i(m-1)-}]$ is considered.

$AW_{i+} = AW_{i+} + Synset[L, AW_{i+}]$ //The function Synset adds to AW_{i+} the synonymous of each word aw_{ij+}, retrieved in the annotated lexicon L. For each synonymous of the word aw_{ij+} new values of co-occurences with the aggregator word rw_{i+} is inserted in PAW_{i+}.

$AW_{i-} = AW_{i-} + Synset[L, AW_{i-}]$ //The function Synset adds to AW_{i-} the synonymous of each word aw_{ij-}, retrieved in the annotated lexicon L. For each synonymous of the word aw_{ij-} new values of co-occurences with the aggregator word rw_{i-} is inserted in PAW_{i-}.

Mining the sentiment from the document

```
for i=0 -> Length[W]

    for k=0 -> Length[$RW_+$]

        if($RW_+$ [k] = = W[i])
            $f_{pre+} = f_{pre+} + 1$

            for(t=0 -> Length[W])
                for(m=0 -> Length[$RW_+$])
                    if($RW_+$ [m] = = W[t])
                        $f_{agg+} = f_{agg+} + prw_{km+}$
                    end if
                end for
            end for

            for(p=0 -> Length[W])
                for(q=0 -> Length[$AW_+$])
                    if($AW_+$ [q] = = W[p])
                        $f_{wag+} = f_{wag+} + paw_{kq+}$
                    end if
                end for
            end for
        end if
    end for

    for k=0 -> Length[$RW_-$]

        if($RW_-$ [k] = = W[i])
            $f_{pre-} = f_{pre-} + 1$

            for(t=0 -> Length[W])
                for(m=0 -> Length[$RW_+$])
                    if($RW_-$ [m] = = W[t])
                        $f_{agg-} = f_{agg-} + prw_{km-}$
                    end if
                end for
            end for

            for(p=0 -> Length[W])
                for(q=0 -> Length[$AW_-$])
                    if($AW_-$ [q] = = W[p])
                        $f_{wag-} = f_{wag-} + paw_{kq-}$
                    end if
                end for
            end for
        end if
    end for

end for
```

Determining the Sentiment

```
$f_p = 0.2\ f_{pre+} + 0.5\ f_{agg+} + 0.3\ f_{wag+}$
$f_n = 0.2\ f_{pre-} + 0.5\ f_{agg-} + 0.3\ f_{wag-}$

if $f_p / f_n > 1.5$
    $Sentiment_D$ = Positive
else
    if $f_n / f_p > 1.5$
        $Sentiment_D$ = Negative
    else
        $Sentiment_D$ = Neutral
    end if
end if
```