IPHV8I3014X
International Journal Of Advanced Research and Innovation -Vol.8, Issue .III
ISSN Online: 2319 – 9253
Print: 2319 – 9245

SENTIMENT ANALYSIS LEVERAGING EMOTIONS AND WORD EMBEDDINGS

[#1]CH. KISHORE KUMAR , *Associate Professor,*
[#2]B.SWAPNA, *Assistant Professor,*
Department of Computer Science,
VAAGDEVI DEGREE AND P.G COLLEGE, HANAMKONDA,WARANGAL, T.S., INDIA.

Abstract: Sentiment analysis and opinion mining are valuable for extraction of useful subjective information out of text documents. These tasks have become of great importance, especially for business and marketing professionals, since online posted products and services reviews impact markets and consumers shifts. This work is motivated by the fact that automating retrieval and detection of sentiments expressed for certain products and services embeds complex processes and pose research challenges, due to the textual phenomena and the language specific expression variations. This paper proposes a fast, flexible, generic methodology for sentiment detection out of textual snippets which express people's opinions in different languages. The proposed methodology adopts a machine learning approach with which textual documents are represented by vectors and are used for training a polarity classification model. Several documents' vector representation approaches have been studied, including lexicon-based, word embedding based and hybrid vectorizations. The competence of these feature representations for the sentiment classification task is assessed through experiments on four datasets containing online user reviews in both Greek and English languages, in order to represent high and weak inflection language groups. The proposed methodology requires minimal computational resources, thus, it might have impact in real world scenarios where limited resources is the case.

Keywords: *Multilingual sentiment analysis ,Text analysis, Machine learning, Vector representation, Hybrid vectorization, Online user reviews.*

I. INTRODUCTION

On a daily basis, millions of people express their views on products, services and offers, among others, using online platforms such as social networks, blogs, wikis, discussion boards, etc. Naturally, the automatic extraction of expressed opinions or implied sentiments in the most accurate manner has become of great importance for businesses, marketing professionals and researchers. Sentiment analysis refers to the process of identifying non trivial, subjective information from a collection of source materials that contain latent information of people's opinions. Such a process can be applied on a variety of textual sources and on different granularity levels ranging from an entire document to individual phrases or, even, separate words. Typically, sentiment analysis reaches characterizations of positive, negative or, sometimes, neutral, for the textual sources at hand.

Sentiment analysis has become an essential part in a wide range of applications and provides benefits for multiple and diverse domains. For instance, sentiment analysis is valuable towards enhancing sales and improving a company's marketing strategies (by tracking customer reviews and survey responses), identifying ideological shifts and analyzing trends in political strategy planning, or, even, forecasting stock market momentum based on world news, financial reports and recorded social media sentiments.

Sentiment analysis algorithms typically employ Natural Language Processing (NLP) processes (such as stemming, part-of-speech tagging, etc.) with utilization of additional resources (e.g. thesauri, sentiment- or emotion-based lexicons, sophisticated dictionaries and ontologies) to model the documents at hand. Important document features are identified towards successful sentiment detection. Such features are, for example: the presence and frequency of terms and parts of speech, opinionated (or emotional) words and phrases, and the existence of negations and intensifiers. Then, a sentiment identification step follows to characterize the textual documents based on their polarity as positive, negative, or neutral. Various techniques can be employed for the sentiment identification step which may be unsupervised or supervised. In unsupervised cases, a *lexicon-based* approach is often used; lexical resources are exploited to assign polarity scores to individual words for detecting the overall sentiment of a document. On the other hand, supervised cases typically follow a *ma- chine learning* approach, where the sentiment detection task is considered as a classification problem by employing algorithms such as Support Vector Machines (SVM), Neural Networks or Naïve Bayes.

Recent and emerging approaches to sentiment detection take advantage of word representations embedded on semantic vector spaces that are learned through the application of neutral networks or probabilistic models on

large text corporation. The derived word embeddings have been shown to accurately capture the semantics and context of words, while their use in a supervised classification setting (especially with neural network architectures) can improve the trained sentiment models.

While lexicon-based approaches result in features that convey information on the overall sentiment orientation of a given document, they often suffer from low coverage, i.e. there are several documents that contain none of the lexicon's words. This is especially evident on short textual snippets, such as those used in online users' communication. Moreover, such approaches fail to capture more latent manifestations of sentiment and emotion, since they do not consider the context in which people ex- press themselves. On the other hand, *word embedding-based* approaches, often employed for constructing vector representations of documents, successfully capture syntactic and semantic regularities encountered in the written language, and there are early results of their beneficial impact on sentiment classification mod- els. However, they do not take advantage of the individual sentiment/emotion value of the words included in a document. In this work, we present a methodology for predicting the sentiment of documents, under the hypothesis that leveraging the strength of lexicons *together* with state-of-the-art word embedding models will result in improved classification performance. There- fore, the proposed methodology derives features at the document-level using: (i) a lexicon-based and (ii) a word embedding-based approach, combined into *hybrid* vectors for a more succinct document representation. The proposed methodology is validated by a series of experiments conducted on four datasets of online user reviews (on movies and technology products, in Greek and English languages). Experiments evaluate the effectiveness of the proposed hybrid vectors in terms of sentiment detection over using separately the lexicon-based or word embedding-based feature vectors. We present a methodology that might be useful for multilingual sentiment analysis, since, in principle, the single language approach restricts the potential and the possible industrial applications of the methodology. The languages under inspection in the current research, English and Greek, have a fundamental difference as far as inflection and morphology are concerned. Modern English, is a typical weak inflection language (e.g. Swedish, Danish), while Greek is a typical high inflection language (e.g. German, Spanish) The distinction between English and Greek, is highlighted in the following short paradigms. In English there is only one form of the adjective *good* while the respective adjective in Greek καλός has 11 different forms (the aforementioned adjective is also related with the sentiment analysis through the sentiment lexicons). In English, there are only 4 forms of the regular verb *ask* (*ask, asks, asked, asking*), while there are 93 different forms of the respective regular verb ρω τ΄αω -Thus, we believe that a satisfactory performance in both languages is encouraging that the methodology could be further applied in other languages.

The main contributions of this work are as follows.

- We present an abstract framework for applying sentiment analysis on different types of textual resources and different languages.
- We propose a hybrid vectorization process that takes advantage of lexicons (with polarized and emotional words) along with state-of-the-art word embedding learning approaches, and demonstrate its effectiveness over an extended set of experiments.
- We test the proposed framework on two Greek (high inflection language) and two English (weak inflection language) datasets (bibliography in the field with respect to the Greek language is still extremely limited).
- We showcase that the proposed sentiment detection frame- work reaches high classification accuracy in all experimentation cases, surpassing existing approaches in literature mainly in terms of efficiency.
- The proposed methodology is suitable for supporting implementation of fast, accurate, flexible multilingual sentiment analysis applications, especially in the context of limited computational resources.

II.RELATED WORK

This section overviews existing research on sentiment analysis, focusing on sentiment detection overall (Section 2.1), with particular emphasis on the Greek language (a typical example of a high inflection language) as a case which imposes specific challenges (Section 2.2).

2.1. Sentiment detection approaches

Sentiment analysis techniques are usually divided into those employing machine learning algorithms operating under a *supervised* setting or statistically inspired methodologies under an *un- supervised* setting. Supervised approaches aim at deriving sentiment classification models, whereas unsupervised methods in- fer the documents' sentiment by exploiting document's statistical properties (in terms of word presence) and/or leverage existing lexicons containing polarized or emotional words. Such lexicons can, however, also be used to derive document representation features that can be used in supervised classification approaches.

2.1.1. Supervised machine learning methods

Sentiment analysis out of a textual resources, can be considered as a document classification problem, aiming at separating documents that express positive and negative sentiments (neutral standing is also considered sometimes) by exploiting certain syn- tactic and linguistic features. Recently, a limited number of more sophisticated approaches have been proposed towards identifying the, more refined, *emotion* of the underlying documents, rather than merely its polarity (i.e. *emotion analysis*). *Emotion analysis*, or *affective analysis* can be considered as a refined version of sentiment analysis, since it aims at a more detailed categorization of documents based on the emotions they express (Chatzakou, Kout- sonikola, Vakali, & Kafetsios, 2013). Both versions of the problem (sentiment and emotion analysis) follow similar approaches, given the existence of a suitably annotated collection of documents. Such annotations can either be nominal (e.g. *positive* , or *anger*) or real valued (e.g. 0.5 *anger* and 0.8 *fear*). Pang, Lee, and Vaithyanathan (2002) are among the pioneers of sentiment analysis employing machine learning techniques to determine whether a written movie review is positive or negative. They experimented with three algorithms for that purpose, i.e., Naïve Bayes, Maxi- mum Entropy and SVM, under a variety of features and parameters, such as unigrams vs. bigrams, and feature frequency vs. feature presence. Socher et al. (2011) proposed a semi-supervised ma- chine learning framework capable of predicting multi-dimensional distributions of the underlying emotion at a sentence-level, based on Recursive Auto-Encoders (RAE). Zhang, Ghosh, Dekhil, Hsu, and Liu (2011) proposed a hybrid technique where an augmented lexicon-based method is first applied to Twitter data to perform entity-level sentiment analysis. Then, a binary classifier receives the results of the preceding step and is trained to assign senti- ment polarities to the opinionated tweets. Cui, Mittal, and Datar (2006) showed the effectiveness of discriminative classifiers, such as SVM, using high order *n* -grams as features for the binary sentiment classification task. Severyn and Moschitti (2015a) discuss the use of a deep Convolution Neural Network for sentiment classification, based on word embeddings that are initialized with the help of a unsupervised neural language model. Ren, Wang, and Ji (2016) utilize Latent Dirichlet Allocation to obtain the topic distribution for each sentence in the dataset and then exploit Recursive Auto-Encoders to learn topic-enhanced word embeddings. To further improve performance, they integrate their representations with traditional models like SVM and logistic regression. Ensemble approaches are proposed by Mesnil, Mikolov, Ranzato, and Bengio (2015) and Wang, Zhang, Sun, Yang, and Larson (2015) . In more de- tail, the work of Mesnil et al. (2015) involves blending both genera- tive (such as Naïve Bayes and Recurrent Neural Networks) and dis- criminative models for sentiment prediction. The log probabilities of these models are combined via linear interpolation to extract the final sentiment assignment, surpassing all competitive models in terms of performance. Wang et al. (2015) propose a new Ran- dom Subspace method, called POS-RS, for sentiment classification based on part-of-speech analysis, which employs both content lex- icon and function lexicon subspace rates to control the diversity of base learners.

2.1.2. Unsupervised methods

These methods follow a fundamentally different approach to sentiment analysis since they do not make use of labelled documents, but rely on documents' statistical properties (e.g. word co- occurrence), NLP processes and existing lexicons with words hav- ing an emotional or polarized orientation. Turney (2002) presents an unsupervised sentiment analysis methodology which classifies reviews as positive or negative, calculating the *semantic orientation* of phrases by associating them with only two words, *excellent* and *poor* . Lin and He (2009) propose an approach that uses La- tent Dirichlet Allocation for detecting a document's sentiment and topic simultaneously, which achieves a classification accuracy similar to that of supervised approaches. There are also unsupervised approaches that rely solely on *lexicons* towards estimating the average sentiment/emotion expressed in the document by the corresponding orientation of its comprising words, according to the lex- icon at hand. Such approaches often start with a small set of opin- ion words with known orientation, and they try to expand that set by utilizing a well known corpus or thesaurus for synonyms. They may also take advantage of structural elements and syntactic patterns that exist in the text by applying NLP processes such as lemmatization and Part-of-Speech (POS) tagging. Heerschop et al. (2011) investigated how knowledge that can be extracted from structural aspects of a document can be utilized to improve the performance of sentiment analysis. They test their hypothesis by identifying the most useful document segments for sentiment detection and score documents based on the aggregation of word-level sentiments. Qiu et al. (2010) extract opinion sentences associated with negative sentiment and find sentence topics, using a rule-based approach that combines syntactic parsing and a sentiment lexicon. They test their approach in contextual advertising, i.e. the problem of associating advertisements with a Web page. Saif, He, Fernandez, and Alani (2016) introduce a lexicon-based approach, called SentiCircles, that is able to update the sentiment orientation of words, by capturing their latent semantics from their co-occurrence patterns. Other works leverage NLP to identify

linguistic features such as negation, intensifiers and modalities, and lexicons to identify the overall sentiment (Carrillo-de Albornoz & Plaza, 2013) or emotion of a document (Chatzakou et al., 2013), under a score averaging approach.

2.2. Sentiment detection on documents written in Greek language

While sentiment analysis of documents written in English has become a very active research area in recent years (as indicated by the variety of approaches discussed above), there has been very little published work on sentiment analysis applications on docu- ments written in the Greek language. However, written Greek is a particularly challenging language for NLP in general, and specif- ically for sentiment analysis, due to its complex morphological features (high inflection, stressing rules, etc.). A relevant effort was presented by Agathangelou, Katakis, Kokkoras, and Ntonas (2014) where an unsupervised iterative approach was employed for mining domain-specific dictionaries from sets of opinionated documents, starting with a small seed of generic opinion words. Their approach is evaluated on a set of user reviews on differ- ent types of electronic and electrical devices from a Greek e-shop. Kermanidis and Maragoudakis (2013) present an unsupervised ap- proach for identifying the sentiment expressed in Twitter posts in order to study the degree of alignment between actual and social web-based political sentiment. They assign simple incidence and frequency values in the posts and build distinct vocabularies for different sets of posts. Solakidis, Vavliakis, and Mitkas (2014) take into account emoticons and lists of emotionally intense keywords in a semi-supervised emotion detection system that is evaluated on a popular Greek student forum. Their system is tested with a number of classification algorithms that include Naïve Bayes, Lo- gistic Regression and SVM, among others.

2.3. Sentiment analysis tools

There is also a growing number of tools, libraries, APIs that can be utilized for sentiment analysis. *Stanford CoreNLP* (Manning et al., 2014) is an integrated framework for performing NLP tasks. It includes a Sentiment Analysis tool that uses deep learning tech- niques and is trained on the *Stanford Sentiment Treebank* , 1 which includes 215,154 phrases, extracted from 11,855 sentences. *Nat- ural Language Toolkit* 2 (NLTK) is a platfrom for building Python programs that utilize human language data. Its sentiment anal- ysis tool, based on text classification, can tag a sentence as be- ing positive, negative or neutral. To achieve that it uses classifiers trained on both twitter sentiment and movie reviews taken from the Movie Review Data. 3 *TextBlob* 4 is a Python library for process- ing input data in the form of texts. It includes an API for common NLP tasks, such as tagging, classification and sentiment analysis. Its sentiment analysis feature can return the polarity and subjectivity score for any given sentence. Other sentiment analysis tools are the *Sentiment API* , 5 *Sentiment140* 6 and *sentimental* . 7 In the following section we present in detail our methodology, the vectorization techniques applied and lexicons used.

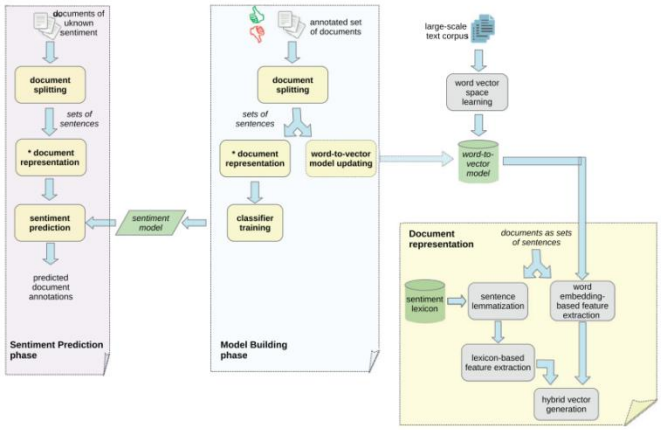

Fig. 1. Sentiment analysis framework: model building and sentiment prediction. ∗*Document representation* process is presented in detail in the bottom-right part of the figure.

III.METHODOLOGY

The methodology of the proposed sentiment analysis frame- work assumes the existence of an annotated collection of documents, which belong to the same domain and are typically polarized (*positive* or *negative*) based on the respective opinions ex- pressed. The proposed methodology is generic since it is not tied to a specific lexicon. It can rather be applied on documents written in various languages, as long as a lexicon that contains polarized or emotion words in this language is available. For a given set of documents, the desired vector representation is extracted, and a classifier is trained to derive a sentiment prediction model. Then, this model can be used to predict the sentiment of new documents of unknown polarity. The proposed methodology is highly customizable since it can function with varying types of vector representations and classification algorithms. As depicted in Fig. 1 the proposed framework supports two main functionalities, which require a specific document representation approach (outlined at the right bottom part of Fig. 1):

- *Model building.* This phase assumes the existence of an annotated collection of documents which will be used for training the sentiment detection model.
- *Sentiment prediction.* This phase assumes the existence of a sentiment model that has been

derived via the *model building* functionality. Given one or more documents, the problem is to predict the conveyed sentiment (per document).

The main advantage of the proposed framework is the extraction of the hybrid feature vectors which, as described above, is a crucial step needed for both the *model building* and *sentiment prediction* functionalities. Therefore, in the following subsections we will provide all the specifics about our vectorization approach. Specifically, we will give the details on the lexicon-based and the word embedding-based feature extraction methods which we utilize, as well as on the proposed hybrid vectors.

3.1. Word embedding-based features

Although lexicon-based features can be used to provide an overall indication of a document's sentiment, they cannot capture more refined characteristics and contextual cues that are inherent in the human language. People often express their emotions and opinions in subtle ways (such as e.g. when they use irony), mix positive with negative polarities or diverse emotions in their expressions, or rely on a set of context-specific expressions/word to communicate their opinion. Recently proposed word embedding- based approaches try to capture semantic and syntactic features of words out of document collections in a language independent pro- cess.

3.2. Word2Vec method

Word2Vec (Mikolov et al., 2013) is an (unsupervised) word embedding-based approach. It aims to detect the meaning and se- mantic relations between words by exploiting the co-occurrence of words in documents belonging to a given corpus. The core idea of Word2Vec is to capture the *context* of words, using machine learning approaches such as Recurrent or Deep Neural Networks.

It actually involves two different learning algorithms: (i) the *Continuous Bag-of-Words* algorithm (CBOW) –whose goal is to predict a word when the surrounding words are given, and (ii) the *Continuous Skip-Gram* algorithm (Skip-gram) – which predicts a set of words when a single word is known. According to Mikolov et al. (2013) , *Skip-gram* operates well with a small amount of training data, representing accurately even rare words or phrases, whereas *CBOW* is much faster to train (than *Skip-gram*) and slightly more accurate for frequent words. Word2Vec operates on a corpus of sentences, first construct- ing a vocabulary based on the words that appear in the corpus more times than a user-defined threshold (to eliminate noise), and then applying either the *CBOW* or the *Skip-gram* algorithm on the input documents to learn the words' vector representations in a *D* -dimensional space. 13 Large textual corpora (e.g. all available Wikipedia articles in a certain language) are often used for training Word2Vec, capturing strong linguistic regularities in the words' relative positions in the learned word vector space that are of a global scope (within the given language.

IV. CONCLUSION

In this work we propose the use of a hybrid approach for the prediction of sentiment, where we combine the context-sensitive coding offered by Word2Vec with sentiment/emotion information offered by a lexicon. The proposed work is based on the hypothesis, that terms' semantic and syntactic relationships, as captured by the Word2Vec representation, or term presence/absence, as captured by a Bag-of-Words representation, are not sufficient for the task of sentiment analysis since they do not carry sentiment in-formation. Thus, the addition of such a lexicon could offer considerable benefits. The resulting hybrid representations are then used as inputs for the supervised training of a classifier which is chosen by the user. The flexibility in the choice of the classifier is assessed with experiments indicating that the SVM model with a linear kernel reaches the best results in terms of efficiency in ac- curacy and the process times (only results with this model have been presented here for brevity). We tested our hypothesis using four different text corpora in Greek and English, along with different coding schemes and different classifier models. The proposed methodology does not surpass the state-of-the- art as far as the accuracy, especially in the English language, is concerned. On the other hand state-of-the-art approaches, most of the time, are computationally costly and their performance is tested only on a single language.

The proposed methodology is simple, fast and flexible and can be applied for any language exhibiting similar properties (customized properly). It provides results of high accuracy in the two languages tested (state-of-the-art for the Greek language). It requires minimal computational resources, thus, it is computationally cheap and might impact in realistic cases. For ex- ample, big data sentiment analysis with limited computational re- sources, or the design of a sentiment analysis standalone software application can highly benefit of the proposed work. In the future we aim to conduct experiments towards the fol- lowing directions, in order to showcase the flexibility and the computational efficiency of the methodology, and further improve its performance: (a) explore new topics beyond online customer re- views, (b) consider another high inflection language and also other language specific phenomena, (c) consider other emotional lexicons for the English language (since Emolex can be attributed with some inefficiencies), (d) apply the methodology into big data

sentiment analysis (e.g. in large social networks datasets such as Twitter textual sources).

REFERENCES

1. Agathangelou, P. , Katakis, I. , Kokkoras, F. , & Ntonas, K. (2014). Mining domain-spe- cific dictionaries of opinion words. In *Proceedings of the web information systems engineering –WISE 2014* (pp. 47–62) .
2. Carrillo-de Albornoz, J. , & Plaza, L. (2013). An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology, 64* (8), 1618–1633 .
3. Cui, H. , Mittal, V. , & Datar, M. (2006). Comparative experiments on sentiment clas- sification for online product reviews. In *Proceedings of the twenty-first national conference on Artificial intelligence (AAAI)* (pp. 1265–1270) .
4. Kotrotsios, K. (2015). Development of tools for the automatic sentiment prediction of greek text using semi-supervised recursive autoencoders. Master's thesis, De- partment of Information Technology, ATEI of Thessaloniki.
5. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint, arXiv: 1301.3781 .
6. Socher, R. , Pennington, J. , Huang, E. H. , Ng, A. Y. , & Manning, C. D. (2011). Semi-su- pervised recursive autoencoders for predicting sentiment distributions. In *Pro- ceedings of the 2011 conference on empirical methods in natural language process- ing* (pp. 151–161) .
7. Severyn, A. , & Moschitti, A. (2015a). Twitter sentiment analysis with deep convo- lutional neural networks. In *Proceedings of the thirty-eighth international ACM SIGIR conference on research and development in information retrieval (SIGIR'15), santiago, chile* (pp. 959–962) .
8. Pang, B. , Lee, L. , & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing: 10* (pp. 79–86) .
9. Taboada, M. , Brooke, J. , Tofiloski, M. , Voll, K. , & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37* (2), 267–307 .
10. Tsakalidis, A. , Papadopoulos, S. , & Kompatsiaris, I. (2014). An ensemble model for cross-domain polarity classification on twitter. In *Proceedings of the Web infor- mation systems engineering –WISE 2014* (pp. 168–177) .

AUTHOR'S PROFILE:

[1]. CH KISHORE KUMAR , presently working as Associate professor at Vaagdevi Degree And P.G College, Hanmakonda, Warangal and Research Scholar(M.Phil) in Bharath University, Chennai. He had 11 years of Teaching Experience, His interested areas are, data mining and data ware house, Information Retrieval System.

[2].B.SWAPNA , presently working as Assistant professor at Vaagdevi Degree And P.G College, Hanmakonda, Warangal. She had 10 years of Teaching Experience, Her interested areas are, data mining and data ware house, Wireless Networks.