Avances en Interacción Humano-Computadora, 2021, Año 6, No. 1, pp. 114-117.
https://doi.org/10.47756/aihc.y6i1.103

A discourse analysis framework for civil action decision-making

Juan Garcia

Published: 30 November 2021

Abstract
Studying the implications of people's opinions on social networks has increased the interest of various stakeholders such as the government, leaders, researchers, and citizens. Consequently, human-computer interaction (HCI) has a vital role through civil action to interact with computational models needed to meet these new demands. By conducting several experiments with a corpus of text data collected from Twitter, we plan to create language representation models based on word embeddings to determine the relevance of discourse concerning a topic and detect abrupt changes over time. Thus, for example, citizens could have quantitative information on the relevance of a political leader's discourse on social issues such as corruption, health, or employment in an electoral process. Alternatively, in a crisis, the authorities could make decisions on the needs of the people by detecting needs expressed in the context changes of the discourse over time.

Keywords:
HCI; Civil AI; Discourse analysis; Word embeddings; Deep learning.

1 Introduction
Opinions posted on platforms like Twitter, when collected over time, can open up possibilities to spot important events that change the course of the conversation. As an example, during the COVID pandemic, it was observed that the term "oxygen" changed its content from nature, parks, forests to health, lack, death due to the crisis experienced in some Latin American countries because of the difficulty of finding oxygen for COVID patients who stayed at home. The analysis of this discourse and its subsequent production of knowledge is an effective tool for decision-making. Thus, citizens can evaluate and make decisions about the proposals of political candidates. Authorities and political leaders can quickly identify social problems.

Garcia, Juan.
Escuela Superior Politécnica del Litoral
Guayaquil, Ecuador
jgarcia@espol.edu.ec.

Universidad de Guayaquil
Guayaquil, Ecuador
juan.garciap1@ug.edu.ec

The advancement of artificial intelligence has dramatically impacted the development of natural language processing (NLP). A feasible approach to analyzing speech is through language models based on vector representations.

The temporal dimension is essential when analyzing the discovery. However, it is crucial to analyze the political discourse in online social networks, discovering nuances beyond positions such as "left-right," "black-white," "good-bad." Other connotations are necessary for analysis and decision-making. Therefore, techniques, metrics, and considerations were applied to understand such nuances in the discussion. We do not have two parties (left and right) in Latin America as in the United States.

2 Related Work
The usefulness of distributed word representations is evident in many natural language processing applications. (NLP) areas, such as part-of-speech (POS) tagging, machine translation, dependency, and semantic parsing, named entity recognition. These distributed representations based on dense vectors, also called word embeddings, can capture similarities and syntactic relationships between words. Typically, vector space representation of words learns from a large unlabeled corpus (e.g., Wikipedia, Giga-word).

This literature review covers different models of semantic representation and their specialization in areas such as policy and techniques for detecting semantic change over time.

2.1 Semantic Representation
Traditional vector space models use linear algebraic techniques in the word co-occurrence count matrix. However, a significant problem with count-based models is that the most frequent words contribute disproportionately to the measure of similarity that negatively impacts analogy tasks. Another approach is to learn word representations that benefit predictions within the windows of the local context. The window is the number of words to predict that can occur in the range of a particular word. A standard method to learning word representations is to train log-bilinear models such as continuous-bag-of-words (CBOW) or the skip-gram architectures as implemented in Word2Vec [11], Glove [13], fast-Text [1]. In the CBOW model, the source word is predicted according to its context, while in the skip-gram model, the predicted nearby words depend on a given source word. Let us assume we are analyzing the following sentence: "*Liberal and progressive spirit to reach an agreement on the fundamental: Peace, Full Democracy and Productive Economy*" as CBOW predicts a word for a given context, e.g., if we use the words' '*liberal*' and '*spirit'* as a context word to predict '*progressive*.' On the other hand, the Skip-gram predicts the context for a particular

word. In the same example, we can give the word 'spirit' as an input to predict the output vectors' liberal' and 'progressive.'

These log-bilinear methods have great success in the application of semantic representations for sentiment analysis [5], political text classification [16], stance detection [19].

Pre-trained representations provide distributional information about words that improve the generalization of learned models over a limited amount of data, such as a small corpus in a specific domain. Usually, the distributional information derives from statistics gathered from a large unlabeled corpus of text data (e.g., Wikipedia). Pre-trained representations can also either be context-free [1] or contextual [9], and contextual representations can further be unidirectional [14] or bidirectional [2]. Context-free models generate a single word embedding representation for each word in the vocabulary; for instance, "right" would have the same representation in "right wing" and "right place." Contextual embeddings capture how words vary across linguistic contexts (i.e., to model polysemy).

2.2 Specialized Word Embeddings

Universal inlays are pre-trained on a large corpus connecting into a variety of downstream task models. These embeddings accurately capture the meaning of common or most frequent words. However, they perform worse on less frequent words [17]. In addition, universal embeddings could overlook crucial aspects of metadata (authors, dates, titles, types). Specialized embeddings can overcome this problem.

Semantic specialization methods generally fall into two categories: a) those that train distributed representations "from scratch" by combining distributive knowledge and lexical information [6]; and b) those that inject lexical information into pre-trained collections of word vectors [10].

Learning from scratch methods modifies the parameters of the embedding algorithm, such as the skip-gram objective function with an emphasis on similarity over relatedness. Although in theory, the word representations produced by the models that train together the baseline information and the local context (discourse) could be as good (or better) as the representations produced by the pre-trained vector. However, their performance has not exceeded that of adjustment methods [12].

Then, the main advantage of the methods that adjust the base vectors is that it provides a portable and lightweight approach to incorporate external or specialized knowledge. In this sense, several works improve the embeddings by adding contextual information of n-grams [6] from specialized lexicons [10]. A recent work evaluated the effectiveness of enriched embeddings from a lexicon of effects applying it to a news dataset. The results show that the use of affect-enriched models significantly improves information retrieval tasks [20].

2.3 Semantic variation

Meaning is not uniform, neither through space nor through time. Across space, different languages such as politic tend to exhibit different polysemic associations for the related terms. In particular, the socio-political discourse has some linguistic characteristics depending on the context, such as the connotation of political language, the set of words about shared agendas, and polysemic. Therefore, in addition to these characteristics, we must add words that can function as a form of social action, which can only be achieved in a specific context.

In meaning across time, technological progress has led to the semantic expansion of terms such as a mouse or apple [4]. In the political arena, the factors leading to a semantic change are more diverse than purely distributive factors [3]. For example, the political discourse can change even within an electoral campaign [15]. In general, there are two recent approaches to the analysis of change of meaning over time. On the one hand, coarse-grained trend analyses compare the semantics of a word in a period with the word's meaning in the preceding period. These coarse-grained models, by themselves, do not specify where a word has changed but only attempt to capture whether a change in meaning has occurred. In contrast, finer-grained [8] analyses usually label the occurrences of words in the corpus and then investigate changes in the corresponding distributions of meaning. These studies indicate that word frequency probably does play a role, but a small one in semantic change [3].

3 Research Questions

This research contains three phases. The first phase is to create a language model as a basis for further discourse analysis. On Twitter, people write a variety of topics, even on a hashtag specific to a topic. The second phase aims to detect content relevant to social-political topics. Finally, the discourse analysis is performed from civil action, generating indicators for decision-making and its evolution over time. The research questions for each phase are following:

- RQ1 How can we assert the accuracy of embeddings?
- RQ2 How to classify political text relevant to a given context (i.e., an event, topic)
- RQ3a How to quantify whether the semantic meaning of discourse addresses or not issues that are relevant for the citizens?
- RQ3b What score beyond the similarity measures the change of the discourse over time?

4 Methodology

A comparative framework can answer the research questions. It covers the context to analyze (horizontal axis) related to the text of the discourse (vertical axis). This framework organizes the discourse in the following three dimensions:

- Time: segment and compare discourse in conventional units such as week or month.
- Participants (political actors, audience): compare the discourse among political actors by quantifying and operationalizing their similarity and divergence.
- Geographical provenance: compare the discourse by geographical area at city and country level.

Learning a specialized socio-political language model is vital and done through semantic specialization, whose effectiveness is analyzed with RQ1. Another important aspect is to remove noise and non-relevant content that may affect the subsequent analysis, where a text classifier is used and analyzed in RQ2. Finally, how to quantify beyond the qualitative is analyzed in RQ3 and its evolution over time.

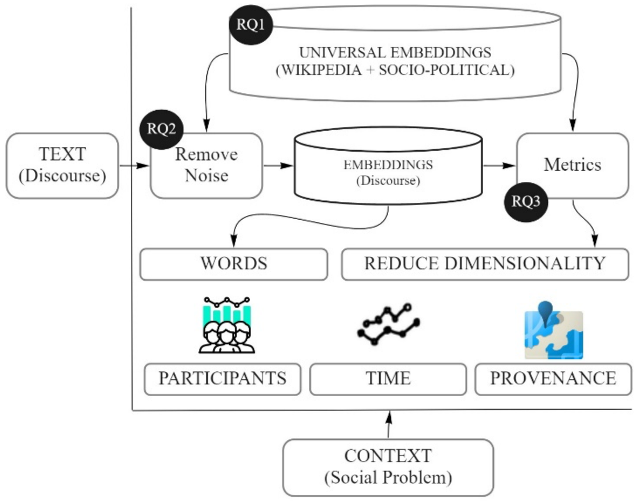

Figure 1. Comparative framework related to Research Questions.

4.1 Data acquisition

Twitter can be a source of information related to socio-political positions expressed in elections, protests, and debates. Public relations between political actors and their different networks (mentions, re-tweet, followings) are valuable information to know how people behave. They are two ways to collect information. First, we can track a political event in real-time using the streaming technique. Second, to analyze past socio-political events, we can use the crawl technique to get tweets starting with a seed of political actors or hashtags.

4.2 Semantic specialization

A pre-trained word distributed representation specialized in the political domain is the basis of the proposed framework. The models analyzed are on the bi-linear techniques word2vec [11], fast-Text [1], and the bidirectional technique BERT [2]. The most recent version available in Wikipedia is used to obtain the universal embeddings that represent the semantic and syntactic knowledge of the words.

Furthermore, it will be done from the socio-political text and the local context to enrich the universal embeddings. For the specialization of the political domain, the training method will be used from scratch [17] with documents taken from multilateral organizations and political science lexicons. In order to enrich the inlays with the local context, a technique of embedding injection [20] will be developed from the local news because this context is dynamic, so training from scratch is costly.

In addition, the evaluation of vector space models performs on word similarity reference points. Spearman's rank correlation serves as the evaluation metric [21]. We include tasks of evaluation of analogies, synonyms, antonyms from the general and specialized domain.

4.3 Identifying a socio-political document

An unsupervised deep learning method classifies socio-political text in two phases. The first phase is to automatically annotate a set of tweets as political and non-political using the political dictionary. Second, to generalize our model, we can train a convolutional neural network (CNN) classifier to distinguish between socio-political relevant and non-relevant. The pre-trained word embedding provides the semantic vector of each word. A mapping process to get the low dimensional vectors takes place before the training and prediction stage.

10-fold cross-validation applied to the training set evaluates the classification method. Next, the F1 score can conduct a variance analysis to observe the impact of the dataset size on the classification performance.

4.4 Context changes over time

Context change detection across periods adopts continuous training by initializing the embeddings for a certain period t with the embeddings trained in the period (t – 1) [7]. The other experiment independently trains an embedding model for each period and then performs a post-hoc alignment [22].

Further, the experiments are compared with two measurements to quantify the context change between two-time points. The first is the cosine distance, and the second measure compares the neighbors of the embeddings.

Two steps systematically evaluate the continuous and independent methods with the cosine and neighboring metrics. First, use synthetic data with three scenarios models with semantic variation, but fundamentally with a four scenario where semantic change does not occur [18]. The second part is with facts. For example, Prada et al. record that candidate Petro made changes in specific contents of the speech during the Colombian elections to achieve greater acceptance, even though he kept the corruption issue as a banner of his campaign.

Finally, a score of relevance ranks the similarities of the socio-political discourse related to a social aspect. Manual analysis with undergraduate and graduate students of communication careers can evaluate the model of whether a word is relevant or not. Spearman's correlation between manual and automatic results can determine the validity of the model.

5 Challenge and Limitations

- Noise: non-relevant content harm the quality of discourse analysis. The socio-political text classifier will include the modeling and feature extraction of multidimensional relations to address this issue.
- Access to data: Due to restrictions on obtaining data from Twitter via public APIs, there is no access to implicit user behavior, such as reading or clicking on a tweet. Because of this limitation, we will focus on explicit behaviors extracted by mining the content of the tweets.

6 Conclusion

Currently, the learning of a specialized language model and the classification of relevant social-political are completed. A universal embedding is crucial since it maintains the generality of the meaning and syntax of the words and amplifies the precision of the metrics and the words obtained according to the social context analyzed. For example, if we analyzed two speeches about proposals to reduce poverty, similar results were obtained compared to an evaluation made with human judgment. Likewise, neural network models are adequate for classifying relevant text with binary classes, for example, if a document is related to a migration context. When classes are multiple, recurrent architectures perform better. We are still working on evolutionary models of semantic change over short periods such as weeks. Previous studies focus on more extended periods such as years or decades. So far, there are indications that the pronounced change of context of a word compared to a previous period is an indicator of important events such as the word "oxygen" in the context of

COVID. Artificial intelligence applied to natural language processing allows the development of this discourse analysis framework under social aspects. The framework generates quantitative information through metrics and qualitative information by producing words that explain the situation, making it an effective tool for decision-making by authorities, political leaders, and citizens.

7 References

[1] Bojanowski, Grave, Joulin, and Mikolov. Enriching Word Vectors with Subword Information. Trans. Assoc. Comput. Linguist. (2017), 135–146.

[2] Devlin, Chang, Lee, and Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL HLT 2019, ACM Press (2019), 4171–4186,

[3] Dubossarsky, Grossman, and Weinshall. Outta control: Laws of semantic change and inherent biases in word representation models. In Proc. EMNLP 2017, 1136–1145.

[4] Eger and Mehler. On the linearity of semantic change: Investigating meaning variation via dynamic graph models. In Proc. ACL 2016, 52–58.

[5] Furini and Montangero. Sentiment analysis and Twitter: a game proposal. In Proc. Pers. Ubiquitous Comput. 2018, 771–785.

[6] Gupta, Pagliardini, and Jaggi. Better word embeddings by disentangling contextual n-gram information. In Proc. NAACL HLT 2019, ACM Press (2019), 933–939.

[7] Hamilton, Leskovec, and Jurafsky. Diachronie word embeddings reveal statistical laws of semantic change. In Proc. ACL 2016, 1489–1501.

[8] Hu, Li, and Liang. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In Proc. ACL 2019, 3899–3908.

[9] Johnson and Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In Proc. NeurIPS 2015, 919–927.

[10] Khosla, Chhaya, and Chawla. Aff2Vec: Affect--Enriched Distributional Word Representations. In Proc. Int. Conf. Comput. Linguist., 2204–2218, 2018.

[11] Mikolov, Yih, and Zweig. Linguistic Regularities in Continuous Space Word Representations. In Proc. NAACL 2013, 746–751.

[12] Mrkšić et al. Semantic Specialization of Distributional Word Vector Spaces using Monolingual and Cross-Lingual Constraints. Trans. Assoc. Comput. Linguist. (2017), 309–324.

[13] Pennington, Socher, and Manning. Glove: Global Vectors for Word Representation. In Proc. EMNLP 2014, 1532–1543.

[14] Peters et al. Deep contextualized word representations. In Proc. NAACL 2018, 2227–2237.

[15] Prada Espinel and Romero Rodríguez. Polarización y demonización en la campaña presidencial de Colombia de 2018. Rev. humanidades 2018, 1–26.

[16] Rao and Spasojevic. Actionable and Political Text Classification using Word Embeddings and LSTM. WISDOM 2016. http://arxiv.org/abs/1607.02501.

[17] Schneider et al. Is Language Modeling Enough? Evaluating Effective Embedding Combinations. In Proc. LREC 2019, 4739–4748.

[18] Shoemark, Liza, Nguyen, Hale, and McGillivray. Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In Proc. EMNLP 2019, 66–76.

[19] Siddiqua, Chy, and Aono. Tweet Stance Detection Using an Attention based Neural Ensemble Model. In Proc. NAACL 2019, 1868–1873.

[20] Teofili and Chhaya. Affect Enriched Word Embeddings for News Information. In Proc. CEUR 2019, 2411.

[21] Vulíc, Glavaš, Mrkšíc, and Korhonen. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. In Proc. NAACL 2018, 516–527.

[22] Yao, Sun, Ding, Rao, and Xiong. Dynamic word embeddings for evolving semantic discovery. In Proc. WSDM 2018, 673–681, 2018.