On the Role of Semantic Word Clusters — CluWords — in Natural Language Processing (NLP) Tasks

Felipe Viegas[1], **Leonardo Rocha**[2] **Marcos André Gonçalves**[1]

[1] Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

[2]Departamento de Ciência da Computação – Universidade Federal de São João del-Rei (UFSJ)

frviegas@dcc.ufmg.br, lcrocha@ufsj.edu.br, mgoncalv@dcc.ufmg.br

Abstract. *The ability to represent data in meaningful and tractable ways is crucial for Natural Language Processing (NLP) applications. This Ph.D. dissertation focused on proposing, designing and evaluating a novel textual document representation that exploits the "best of two worlds": efficient and effective frequentist information (TFIDF representations) with semantic information derived from word embedding representations. In more details, our proposal – called* ***CluWords*** *– groups syntactically and semantically related words into clusters and applies domain-specific and application-oriented filtering and weighting schemes over them to build powerful document representations especially tuned for the task in hand. We apply our novel Cluword concept to four NLP applications: topic modeling, hierarchical topic modeling, sentiment lexicon building and sentiment analysis. Some of the novel contributions of this dissertation include: (i) the introduction of a new data representation composed of three general steps (clustering, filtering, and weighting). These steps are specially designed to overcome task-specific challenges related to noise and lack of information; (ii) the design of CluWords' components capable of improving the effectiveness of Topic Modeling, Hierarchical Topic Modeling and Sentiment Analysis applications; (iii) the proposal of two new topic quality metrics to assess the topical quality of the hierarchical structures. Our extensive experimentation demonstrates that CluWords produce the current state-of-the-art topic modeling and hierarchical topic modeling. For sentiment analysis, our experiments show that CluWords filtering and weighting can mitigate semantic noise, surpassing powerful Transformer architectures in the task. All code and datasets produced in this dissertation are available for replication. Our results were published in some of the most important conferences in journals of the field, as detailed in this document. Our work was supported by two Google Research Awards.*

1. Introduction, Motivation and Problem Defition

In this Ph.D. dissertation, we delve into the realm of text representation, focusing particularly on word embedding models such as Word2Vec [Mikolov et al. 2013], GloVe [Pennington et al. 2014] and FastText [Mikolov et al. 2018]. Static Word embeddings generate single representations for words based on co-occurrences and contextual information, encapsulating semantic and positional relatedness in high-dimensional vectors. While contextual word embeddings, exemplified by transformer architectures such as BERT [Devlin et al. 2018], excel in capturing word meanings within contexts, recent studies [Dufter et al. 2021] have shown the efficacy of static word embeddings, such as FastText, in several contexts and tasks such as document Classification. The superiority of embedding models over count models like TF-IDF has been consistently demonstrated across various tasks, as shown by Baroni et al. [Baroni et al. 2014]. Our proposed solution harnesses the power of static embedding representations to enrich data representation without being constrained to any specific embedding model[1].

Texts have been traditionally represented using fixed-length vector representations, commonly known as bag-of-words (BoW), where each vector's length equals the collection's vocabulary size. Despite its conceptual simplicity and computational efficiency, BoW suffers from sparsity issues in collections where documents contain only a fraction of the vocabulary. Strategies like n-grams and compound features have been adopted to capture such semantic relationships and reduce ambiguity and noise in BoW representations. Recent efforts in Natural Language Processing (NLP) have witnessed the adoption of static word embeddings to enrich data representation, exploiting the inherent semantic relationships encoded within word embeddings.

Though richer than previous models, embedding models, in general, may still suffer from semantic noise, in which words with different meanings may have high similarity in the vector space. Semantic noise happens mainly in two scenarios: (i) the application domain is distinct from the domain in which the embeddings were created. (ii) the set of training documents used to fine-tune the embedding vector space is small. In both scenarios, the learned embedding model may not capture the correct information about the words, especially regarding infrequent words [Nooralahzadeh et al. 2018], leading to semantic noise due to the lack of information.

In this context, we main contribution of this PhD dissertation is a new concept called **Cluwords** – a novel text representation that leverages the efficiency and interpretability of Bow-based representations while taking advantage of the richer semantic capablities of word embeddings. In more detail, our CluWords framework comprises three fundamental steps —- clustering, filtering, and weighting – aimed at constructing a more informative representation for textual data collections tailored to specific application scenarios. CluWords entail clusters of semantically related word embeddings [Mikolov et al. 2018], formed through the application of distance functions and customizable filtering mechanisms. Our main goal with CluWords lies in leveraging word embedding similarities, particularly by filtering out noise and appropriately weighting terms, to construct enriched and adaptable word representations.

Our main *hypothesis* is that *Cluwords are a better alternative (more effective) for representing text, especially in small, noisier datasets suffering from information scarcity and noise, as they adeptly capture semantic relationships along with frequentist information, crucial for several NLP tasks.* We empirically demonstrate our main hypotheses through extensive experimentation deploying CluWords in three application scenarios: Topic Modeling (TM) [Viegas et al. 2018, Viegas et al. 2019], Hierarchical Topic Modeling (HTM) [Viegas et al. 2020b], and Sentiment Analysis [Viegas et al. 2020a, Viegas et al. 2023], showcasing its effectiveness in enhancing various NLP tasks.

[1] Although our initial focus was on static word embeddings, the dissertation has experiments and discussions on the issues of using contextual word embeddings in our proposals.

2. Summary of the Contributions

Our main general research questions, which are derived from our main hypothesis, are: *(i) Can CluWords be effectively exploited to advance the state-of-the-art in NLP and Information Retrieval tasks? (ii) Are task-specific filtering and weighting effective mechanisms capable of adapting the CluWords to different NLP application scenarios?*. The specific research questions that may support our general research questions, considering three application scenarios, include:

1. **Topic Modeling:** *(i) Can we exploit the CluWords to enhance document representation for topic modeling? (ii) Can the CluWords add more information to hierarchical topic modeling models at deeper levels of the hierarchy?*
2. **Sentiment Analysis:** *(i) Can the CluWords be used to overcome issues of lack of information in sentiment analysis tasks? (ii) Can polarity/intensity and Part-of-Speech (PoS) be used to filter out words from CluWords for sentiment analysis?*

Our experimental findings provide compelling evidence towards positively answering the first research question in the TM context. Through experiments using 12 datasets and eight baselines, we confirm that CluWords indeed can build better topics and significantly enrich document representations.

Delving deeper into the second research question, dealing with HTM, we introduce a novel unsupervised non-probabilistic method called CluHTM. This method exploits the global semantic information provided by the CluWords representation and an original application of a stability measure to define the hierarchy "shape". We present two variants of the CluHTM method – (i) f-CluHTM, which exploits pre-trained static embeddings to build the CluWords representations, and (ii) c-CluHTM, which exploits the Contextual word embeddings from BERT-transformed into static version using pooling approaches to build the CluWords representations. CluHTM variants excelled, being around twice as effective as the strongest state-of-the-art baselines.

In addition, we present new evaluation metrics for evaluating HTM methods. The newly proposed topic quality metrics assess aspects related to topological consistency (or redundancy) and the hierarchical semantic structure that are important to hierarchical methods. These are different and complementary aspects than those captured by traditional TM metrics such as NPMI and Coherence, which do not consider topological relationships among topics. Our new proposed topic quality metrics capture distinct behaviors from the topics built, including duplicity and semantic consistency. Our results also show that c-CluHTM and f-CluHTM present the best results in building a hierarchical structure while avoiding redundancy.

Transitioning to the domain of Sentiment Analysis (SA), regarding research questions RQ2.i and 2.ii, firstly we provided formal hypotheses supported by strong empirical and experimental evidence demonstrating the potential of exploring CluWords in SA. In addition, we proposed a new, simple, yet very effective technique for expanding human-built lexicons. Our method can use the general representation of words provided by word embeddings and their relationships (captured by simple distance computations) to produce **high coverage** lexicons that significantly improve accuracy. Secondly, we presented a new instantiation of the CluWords for SA – CluSent – that exploits semantic expansion and tackles information shortage and noise issues. CluSent representation is built by a dynamic pipeline of instantiations to build dataset-oriented document representations. Our experimental evaluation reveals that CluSent, through PoS-based filtering and sentiment weighting, is as effective as the best Transformer methods for the SA task.

Our dissertation granted two Google Latin American Research Awards (LARA) – eighth and ninth editions. In addition, our work on building and applying the concept of CluWords in NLP tasks has been validated and published in the main Information Retrieval and Natural Language Processing conferences and journals in the past years, including: (1) ACM Conference on Information and Knowledge Management 2018 (Qualis A1); (2) ACM International Conference on Web Search and Data Mining 2019 (Qualis A1); (3) AAAI Annual Meeting of the Association for Computational Linguistics 2020 (Qualis A1); (4) Information Systems Journal, 2020 (Qualis A2); (5) Scientometrics Journal, 2022 (Qualis A1) (6) WebMedia 2023 (Qualis A4); (7) Computational Linguistic, 2024, 2nd round of review; (8) Journal on Interactive System, 2024 (Qualis B1, (submitted)). We refer the reader to the companion "Sub-products" document for a detailing of these publications.

Indirect contributions, inspired or influenced by our work, can be found in papers, published in the following venues: (1) ACM International World Wide Web Conference 2018 (Qualis A1); (2) International Conference on Computational Science and Its Applications 2021 (Qualis A3); (3) Information Processing & Management Journal, 2021 (Qualis A1); (4) Webmedia 2023 (Qualis A4); (5) Information Processing & Management Journal, 2023 (Qualis A1); (6) ACM Computing Surveys, 2023 (Qualis A1). We also refer the reader for the companion document for a full detailing of these publications.

3. Development – The CluWords Framework

Our main motivation behind CluWords is to construct document representations that leverage word embedding models, to capture semantic relationships among words, together with frequentist information, thereby enhancing the informational content of documents. However, this task is complicated due to semantic noise inherent to embedding models, where words with distinct meanings may exhibit high similarity in vector space. We introduce CluWords, clusters of semantically related words integrated with task-specific filtering and weighting schemes to address this challenge. CluWords enhance data representation by leveraging word embedding similarities, filtering out semantic noise, and appropriately weighting terms based on the application context. The CluWords framework encompasses three key steps: (i) clustering; (ii) filtering; and (iii) and weighting. Clustering involves strategies to capture semantic relationships between words using embedding models, potentially utilizing distance-based metrics and classic clustering methods. Filtering mechanisms are then applied to filter out noise from the generated clusters. Finally, weighting combines semantic information with statistical information, employing strategies to weight the relevance of terms based on syntactic information within documents and the textual collection.

3.1. Instantiating CluWords for Topic Modeling

This Section presents how we instantiate the CluWords steps for the TM tasks. Figure 1.A presents the process of transforming each original word into a CluWord representation, described as follows: (i) The clustering and filtering steps exploit the nearest neighborhood approach based on the information about the dataset, combined with word

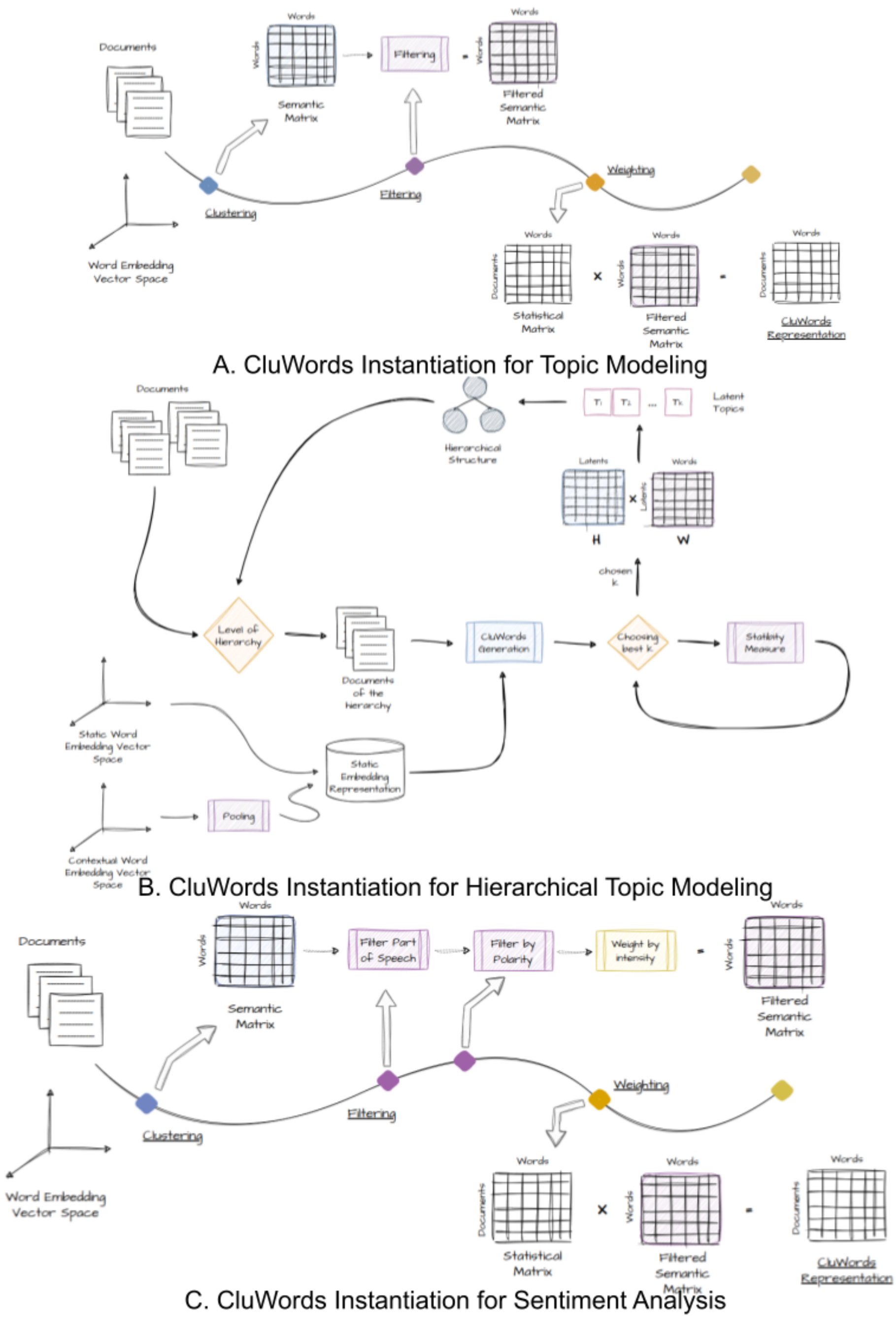


Figure 1. **Cluwords instantiations for the NLP domains explored in the dissertation.**

embedding representation; (ii) The weighting step exploits both semantic and statistical information combined to measure the importance of each CluWord in a modified version of the TF-IDF weighting scheme.

Clustering and Filtering The CluWords representation is constructed using a semantic matrix, denoted as C, where each dimension corresponds to the vocabulary size ($|\mathcal{V}|$). The semantic matrix C is defined by computing cosine similarities between word vectors, represented as u, belonging to the dataset vocabulary $\mathcal{V}$. Each entry $C_{t',t}$ of C represents the similarity between words t' and t, computed using a cosine similarity function $\omega(u_{t'}, u_t)$. A similarity threshold α is employed as a regularizer, with larger values of α leading to sparser representations. Entries of C exceeding α are retained as non-zero values, forming CluWords, while others are set to zero. This formulation effectively creates a compact representation of semantic relationships between words in the dataset vocabulary, facilitating subsequent analysis and processing.

Weighting In Figure 1.A, the CluWords representation combines the statistical (term-frequency) and semantic C matrices. The statistical matrix (TF) denotes word frequencies in documents ($TF \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{V}|}$), with $TF_{d,t}$ representing the word t frequency in document d. For a CluWord (CW) t in a document d, its representation is the product of $\overrightarrow{TF_d}$ (document term-frequencies) and $\overrightarrow{C_{,t}}$ (semantic scores). Additionally, the TF-IDF weighting for a CluWord t in document d is obtained by multiplying its representation by idf_t, the inverse document frequency component, calculated as $log\left(\frac{|\mathcal{D}|}{\sum_{1\leq d\leq|\mathcal{D}|}\mu_{t,d}}\right)$. The average $\mu_{t,d}$ is determined as the mean of semantic weights for CluWord t across terms in document d where $C_{t',t}$ is non-zero in $\overrightarrow{C_{,t}}$.

3.2. Instantiating CluWords for Hierarchical TM: CluHTM

The CluHTM method explores a similar configuration to the CluWords. Its intuition consists of smoothing the limitation of non-probabilistic methods in exploring local information by using CluWords to explore *global* semantic information into the hierarchical topic tree creation. To automatically generate the hierarchal structure, as shown in Figure 1.B, we exploit the Stability Measure. The CluWords Generation block in Figure 1.B is similar to the instantiation shown in Figure 1.A.

Stability [Greene et al. 2014] aims to assess the consistency of topics by evaluating the overlap of top words extracted from multiple random samplings. Given a range of topics and a TM method, such as the Non-negative Factorization Matrix, the process involves first learning a topic model on the complete dataset to establish a reference point for topic word distributions. Subsequently, multiple random subsamples are drawn from the dataset, generating a topic model for each subsample. Stability quantifies the agreement between the top words of the reference model and those of each subsampled model.

3.3. Instantiating CluWords for Sentiment Analysis: CluSent

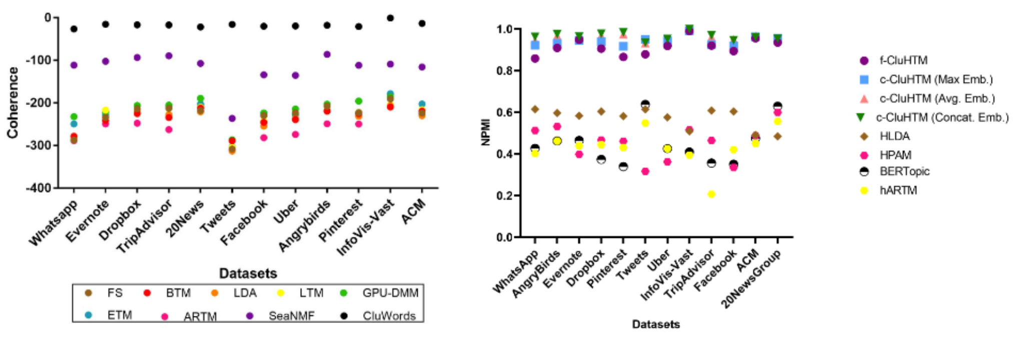
Inspired by our previous work that showed that word embedding may carry sentiment information [Viegas et al. 2020a], we designed CluSent for Sentiment Analysis (SA).

CluSent applies the same three generic steps to a given source text representation: clustering, filtering, and weighting to build a richer (more informative) representation for a textual collection. The main differences are in the two latter steps.

Figure 1.C illustrates how CluSent representations are instantiated for a given collection. As before, CluSent groups semantically related word embeddings (first blue dot in Figure 1.C) Filtering comprises two sub-steps. The first filter is the **Part-of-Speech Filter** – the intuition is that, for SA sake, we only want to keep adjectives that are semantically similar, the same for adverbs, and so on. Thus, we exploit a Part-of-Speech tagging to filter out words from the neighborhood of a CluWord w. The second instantiation block is the **Sentiment Filtering and Weighting**. It is a joint filtering and weighting block that exploits a lexicon dictionary (expanded version of the VADER [Hutto and Gilbert 2014]) to filter out words with different polarities from the neighborhood of a CluWord w. In addition, this block also uses the sentiment score of the lexicons to weight the Semantic Matrix C. The remainder steps are similar as before.

4. Experimental Results

In the context of Topic Modeling, we exploit the CluWords representation to enhance non-probabilistic TM [Viegas et al. 2018, Viegas et al. 2019]. One of the main CluWords advantages is that they can be used in standard BoW representations, just as the original words, along with standard algorithms, without having to resort to sophisticated solutions such as pooling. The key ideas that make CluWords work so well in this application are a dynamic threshold that controls the neighborhood size (to control noise) and a new weighting scheme specially designed to weight this type of evidence in a BoW-like representation. Our results are currently state-of-the-art in TM, with gains of 80% over the best baseline in our experiments. Figure 2.A illustrates one of the experimental results performed in the experimental evaluation, contrasting the Coherence scores considering the top 10 words for the topics.



A. CluWords B. CluHTM

Figure 2. Comparing the results achieved in the Topic Modeling and Hierarchical Topic Modeling experiments, respectively. Figure A is the results achieved by each strategy considering the top 10 words for TFIDF-Coherence in the TM scenario, while Figure B NPMI results considering the top 10 words in the Hierarchical TM scenario.

Regarding the HTM context, one of the main advantages of CluHTM is that it exploits a cross-level stability analysis metric for defining the number of topics and ultimately 'the shape' of the hierarchical structure in an original way; as far as we know *this metric has never been applied with this goal*. In our experimental evaluation, some of our results improved over the state-of-the-art by more than 500%. We introduced two CluHTM variants, the f-CluHTM, which uses static embeddings in the CluWords generation, and c-CluHTM, which exploits pooled transformations of BERT's contextual embeddings. This c-CluHTM variant exploits the BERT's hidden layers by exploiting several pooling strategies (e.g., average, maximum, concatenation combined with pooling) to build a single-word embedding representation for each word in CluHTM's meta-word construction. The idea is to evaluate the quality of the topics built with these new word embedding representations in the CluHTM solution compared to the f-CluHTM solution. Our experimental results showed that the CluHTM variants can improve the document representation, generating more cohesive topics in all evaluated hierarchy levels. Our experimental evaluation demonstrates the superiority of c-CluHTM with gains between 12% and 21% regarding NPMI (Figure 2.B) and Coherence. In addition to the CluHTM proposal, we observed a lack of metrics to evaluate the quality of topics across the hierarchy structure. So, we designed two new topic quality metrics to assess topical quality, considering two aspects: (i) *Topic topological consistency (or redundancy)* and (ii) *Semantic hierarchical structure*. Accordingly, we proposed Uniqueness and Semantic Hierarchical Structure (SHS) metrics, which can capture redundancy and semantic relationships over the topics of the hierarchical structure. Regarding these two metrics, the CluHTM variants can achieve gains of up to 60% compared to the same (best) baselines.

Finally, in the context of **Sentiment Analysis**, we challenged the notion popularized in [Tang et al. 2014] that we should not exploit (general) semantic relationships based on vector proximity by expanding existing manual lexicons with unknown words. The affective (sentiment) values of these new words (previously unknown in the lexicon) can be automatically derived from the neighborhood of word embeddings in the lexicon. We performed empirical experiments to show that semantic relationships can be effective and exploited for building Sentiment Lexicons. Thus, we proposed a new, simple, yet very effective technique for expanding human-built lexicons. Our method can use the general representation of words provided by word embeddings and their relationships (captured by simple distance computations) to produce **high coverage** lexicons that significantly improve accuracy. Using the extended lexicon generated with our method to predict the sentiment of sentences, we achieved gains in MacroF1 of up to 26% and 38% against RoWE and SENTPRO, two of the best baselines.

Inspired by the previous semantic analyses, we proposed CluSent, adding new filtering and weighting instantiations using the sentiment lexicon's polarity and intensity to tackle the problems of information shortage and noise, commonly found in sentiment analysis applications. The CluSent proposal is a *demonstration* of how to build and dynamically instantiate the CluWords' filtering (aiming at de-noising) and weighting mechanisms by exploring polarity and intensity information from unsupervised lexicons. In our experimental evaluation, comparing *CluSent* with five strong state-of-the-art sentiment analysis baselines in a large benchmark with 19 datasets, our solution achieved the best results in 30 out 38 possibilities (19 datasets considering MacroF1 and MicroF1), with gains up to 14.21% (*ss_bbc*), 7.60% (*ss_digg*) and 7.17% (*ss_rw*) compared to the *best baseline in each dataset*, in terms of MacroF1 (Table 1). In addition, CluSent outperformed the BERT Transformer in 12 out of 19 datasets, tying in other 6 and losing only in 2 cases.

Dataset	BERT	NB-W-B + dv-cos	RNTN	L-MIXED	kNN Reg. Exp.	CluSent
aisopos_tw	**86.73**	**84.74**	63.63	83.58	82.95	**87.74**●
debate	73.79	66.42	62.4	**77.41**	61.53	**75.13**●
narr_tw	79.71	63.42	74.12	**82.48**	**83.46**	**86.50**●
pappas_ted	73.52	74.85	63.42	**77.64**	65.43	**78.82**●
sanders	**78.07**	76.29	68.02	**80.47**	69.81	**80.37**●
ss_bbc	55.99	46.48	55.55	51.28	60.36	**68.94**▲
ss_digg	65.68	43.20	66.05	55.87	65.55	**71.07**▲
ss_myspace	61.02	45.67	62.47	49.88	**75.35**	**73.35**●
ss_rw	70.56	42.12	62.90	57.72	67.53	**75.62**▲
ss_twitter	72.21	55.99	68.17	**74.81**	**73.94**	**75.44**●
ss_youtube	76.55	54.40	71.31	**79.69**	**77.09**	**79.02**●
stanford_tw	75.70	72.88	77.52	**79.54**	**81.41**	77.07▼
semeval_tw	**74.09**	48.60	68.92	68.37	**75.52**	**76.51**●
vader_amzn	**71.48**	62.85	69.33	**73.89**	62.49	**71.94**●
vader_movie	78.09	76.59	75.31	**82.63**	64.59	75.11▼
vader_nyt	**65.56**	53.19	60.92	**66.92**	**66.00**	**65.56**●
vader_tw	81.92	61.23	71.67	82.53	**89.25**	**89.63**●
yelp_review	**94.08**	**93.30**	74.33	**94.59**	62.46	**92.36**●
SST-2	**94.39**	86.87	82.75	**93.13**	55.11	89.02 ▼

Table 1. MacroF1 results. CluSent is the best method (winning or tying) in 16 out of 19 datasets.

5. Conclusions and Future Work

In this PhD dissertation, we introduced a a novel textual representation that exploits the "best of two worlds": frequentist information, similar to what is done in efficient and effective Bow-based representations, along with semantic information derived from word embeddings. Our **CluWords** groups syntactically and semantically related words into clusters and apply domain-specific and application-oriented filtering and weighting mechanisms over them to build powerful document representations especially tuned for the task at hand. Our extensive experimental evaluation demonstrated that the employment of Cluwords not only produces the current state-of-the-art (unbeaten so far) in Topic Modeling tasks, including Hiearchical ones, but generates better results than powerful Transformers in sentiment analysis. Perhaps even more important, the application of the same concept in different tasks shows the flexibility and adaptability of our proposals. Cluwords is a novel and, in our opinion, a revolutionary proposal that may have a very positive impact in many NLP applications, as it has a lot of expressive power, capturing semantics as word embeddings, but with a much more efficient and interpretable BoW-style implementation.

We envision many possibilities of extensions for the works presented in this dissertation. In the context of TM and HTM, we will test new Cluword instantiations, e.g., with filters similar to those designed for CluSent and new embedding representations, especially multi-lingual embedding representations. We believe that CluWords has the potential to deal with multi-language domains. We will also investigate new ways of evaluating HTM metrics. We believe that existing metrics, such as those used for hierarchical clusters, could potentially be used for evaluating HTM strategies.

In the context of sentiment analysis, we believe that further exploiting contextual embedding representations in different ways may help improve the CluWords representation for this task. We also believe that CluWords has the potential to be exploited in other domains, such as recommendation systems and search systems. Regarding Recommendation Systems, there is plenty of matrix-based strategies, such as Factorization Machines and Review-Aware Recommender Systems (RARDs) that can be easily exploit CluWords. Regarding search systems, we believe that CluWords could be used as a method for query expansion to improve the quality of queries.

Acknowledgments

This Ph.D. thesis was funded by Google, AWS, CNPq, CAPES, FINEP, and Fapemig.

References

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL'14*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dufter, P., Kassner, N., and Schütze, H. (2021). Static embeddings as efficient knowledge bases?

Greene, D., O'Callaghan, D., and Cunningham, P. (2014). How many topics? stability analysis for topic models. *CoRR*.

Hutto, C. J. and Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM'14*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in pre-training distributed word representations. In *LREC'18*.

Nooralahzadeh, F., Øvrelid, L., and Lønning, J. T. (2018). Evaluation of Domain-specific Word Embeddings using Knowledge Resources. In *LREC 2018*.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL 2014*, pages 1555–1565.

Viegas, F., Alvim, M. S., Canuto, S., Rosa, T., Gonçalves, M. A., and Rocha, L. (2020a). Exploiting semantic relationships for unsupervised expansion of sentiment lexicons. *Information Systems*, 94:101606.

Viegas, F., Canuto, S., Cunha, W., França, C., Valiense, C., Rocha, L., and Gonçalves, M. A. (2023). Clusent – combining semantic expansion and de-noising for dataset-oriented sentiment analysis of short texts. In *Webmedia 2023*, page 110–118.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. In *WSDM '19*.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., and Gonçalves, M. A. (2020b). Cluhtm: Semantic hierarchical topic modeling based on cluwords. In *ACL'2020*.

Viegas, F., Luiz, W., Gomes, C., Khatibi, A., Canuto, S., Mourão, F., Salles, T., Rocha, L., and Gonçalves, M. A. (2018). Semantically-enhanced topic modeling. In *CIKM '18*.