UNIVERSIDAD DE LAS AMÉRICAS PUEBLA

Escuela de Artes y Humanidades

Departamento de Lenguas

Contributions to Social Learning Analytics based on Sentiment Analysis of Students' Interactions in Educational Environments

Tesis que, para completar los requisitos del Programa de Honores presenta la estudiante

María José Díaz Torres

ID: 153452

Licenciatura en Idiomas

Dra. Ofelia Delfina Cervantes Villagómez

San Andrés Cholula, Puebla. **Primavera 2019**

Contributions to Social Learning Analytics based on Sentiment Analysis of Students' Interactions in Educational Environments

Abstract

This study describes a sentiment analysis service that is part of a learning analytics platform developed for the Uruguayan educational system, and proposes four new localized sentiment classification models. The sentiment analysis service performs the natural language processing task of determining the attitude or sentiment associated to a text, in this case, the sentiments of student-generated comments as a result of their interactions in several learning management systems and social media. The methodology of the original sentiment classifier is discussed and the proposal of possible improvements to the system is made from a linguistic perspective. The proposal consists in adapting the generic Spanish classifier, based on an international Spanish corpus, to create a localized Uruguayan (Rioplatense) Spanish sentiment classifier. This process involves enriching the model with regional vocabulary and expressions, training the system in a dialect-specific dataset and using a number of text representation features, including n-grams, POS tags, and a variety of stylistic features. To build the models different machine learning algorithms were used, such as SVM, Naïve Bayes, logistic regression and a decision tree. The results of the testing reveal that the all of the four proposed localization approaches outperformed the original sentiment classification model.

Keywords: *linguistic variation, machine learning, Rioplatense Spanish, sentiment analysis, social learning analytics, Uruguay.*

Resumen

Este estudio describe un servicio de análisis de sentimientos, que forma parte de una plataforma de analítica del aprendizaje desarrollada para el sistema educativo uruguayo, y propone cuatro nuevos modelos localizados de clasificación de sentimientos. El servicio de análisis de sentimientos realiza la tarea de procesamiento de lenguaje natural de determinar la actitud o sentimiento asociado a un texto, en este caso, los sentimientos de los comentarios generados por los estudiantes como resultado de sus interacciones en varios sistemas de gestión de aprendizaje y redes sociales. Se discute la metodología del clasificador de sentimientos original y se realiza una propuesta de posibles mejoras al sistema desde una perspectiva lingüística. La propuesta consiste en adaptar el clasificador de español genérico, basado en un corpus de español internacional, para crear un clasificador de sentimiento de español uruguayo (Rioplatense) localizado. Este proceso implica enriquecer el modelo con vocabulario y expresiones regionales, entrenar al sistema en un conjunto de datos específico del dialecto y usar diferentes representaciones textuales, incluyendo n-gramas, categorías gramaticales y una variedad de rasgos estilísticos. Para construir los modelos se utilizó una serie de algoritmos de aprendizaje automático, como SVM, Naïve Bayes, regresión logística y un árbol de decisión. Los resultados de las pruebas revelan que los cuatro enfoques localizados propuestos superaron al modelo de clasificación de sentimiento original.

Palabras clave: *análisis de aprendizaje social, análisis de sentimientos, aprendizaje automático, español rioplatense, Uruguay, variación lingüística.*

Agradecimientos

Quiero agradecer a mi familia, a mis padres por compartirme su experiencia y perspectiva, por ser mis guías, amorosos y pacientes. Les debo todo y nunca dejaré de agradecerles. Gracias, mamá. Gracias, papá.

A mis hermanas, por su cariño, preocupación y curiosidad sobre mi investigación, pero también por estar ahí para olvidarme de ella un rato y divertirnos. Gracias por aguantarme y apoyarme siempre.

A Rodrigo, mi compañero y mejor amigo, gracias por tu amor y apoyo incondicional, por recordarme que siempre debo orientar mis decisiones hacia mi felicidad. Sin ti no habría llegado hasta donde estoy.

A mis amigas y amigos, las maravillosas personas que conocí en la universidad y también a las que volví a encontrar en el camino. Gracias por todas las risas, abrazos y palabras de aliento, por estar en los mejores y los peores momentos.

A mis profesores, por siempre esforzarse en dar más de ellos y mostrar lo que es la pasión por lo que haces. Gracias por responder cada pregunta, por apoyar mi curiosidad, y por tener siempre sus puertas abiertas.

Index

Index of Tables

Index of Figures

1. Introduction

This research study fits within the larger framework of *Plan Ceibal*[1], a socio-educational project of Uruguay created in 2007 to support educational policies aimed at digital inclusion and equal opportunities through technology (Plan Ceibal, 2017). The main goal of the program is to narrow the digital gap not only in comparison with other countries but also within Uruguay itself (Plan Ceibal, 2017). The plan undertakes the responsibility for conducting educational research, evaluating and constantly training educators, and creating programs and resources to achieve its goals. Accordingly, the main action of the plan was to provide every student and teacher in the public education system at the national level with a portable computer for their personal use with free Internet connection at their educational institutions (Plan Ceibal, 2017). Further, Plan Ceibal provides educational support services by means of two learning management systems (LMSs): the educational social network "CREA 2" and the Adaptive Mathematic Platform "PAM" (Plan Ceibal, 2017).

Despite Plan Ceibal's efforts, it was later shown that access to technology does not ensure the fulfillment of the main objectives by itself; the distribution of laptops did not influence the Uruguayan students' academic performance (De Melo, Machado, Miranda & Viera, 2013). Moreover, teachers do not seem to have fully integrated them as resources to improve learning but rather as tools for information search (Fullan, Watson & Anderson, 2013), in spite of the availability of the LMSs. In this respect, educational proposals to

[1] www.ceibal.edu.uy

leverage the new technologies are needed, to develop strategies that foster educational leaders' empowerment and the innovative use of the existing technology and infrastructure.

Following from this need, we proposed the DIIA project. DIIA was an international collaborative proposal was put forth by the University of the Republic (UdelaR, Uruguay), the University of the Americas (UDLAP, Mexico), the *Consejo de Formación en Educación* (CFE, Uruguay) and the *Centro Regional de Profesores del Suroeste* (CeRP, Uruguay) with financing from the National Agency for Research and Innovation of Uruguay (ANII) 2016 fund for digital inclusion.

The DIIA project (Discovery of Interactions that Impact in Learning) involves the development of a software service for the discovery of semantic patterns that have an impact on learning, based on students' interaction in social learning networks. This proposal is based on social learning analytics, and thus includes the analysis of interactions that occur in social settings, not only students-materials interactions but also student-student and student-teacher's interactions. In this sense, through the DIIA project we argued that these patterns convey critical information for decision-making at both classroom and institutional planning level, since they support teachers and educational agents in making strategic decisions to improve the learning experience of students. To meet these ends, the DIIA team proposed a visualization platform providing strategic analytical data about student performance and interaction in both institutional and informal learning platforms, namely CREA 2 and Facebook. The DIIA software service incorporates different pattern detection algorithms with approaches that consider the semantic nature of social interactions and the participation of students in multiplatform contexts, hence allowing for social learning analytics.

One innovative approach to social learning analytics adopted by DIIA was the development of a sentiment classifier, a natural language processing application that determines the polarity of the attitude or sentiment of a writer with respect to some topic (Pang & Lee, 2009). In other words, this component would take students' short texts generated in the learning platforms, such as comments, posts and messages, and assign them a positive, negative or neutral polarity. Hence, this service provides meaningful insights into students' opinions about the courses in general, workload, materials, and, moreover, about their emotional state and interpersonal relationships with other students and teachers. Regarding the methodology, the sentiment classifier is based in a supervised learning model and predicts the polarity of the texts based on lexical-syntactic sequential elements, features that characterize the documents of each class (positive, negative and neutral messages). The model was trained with the InterTASS-2017 corpus, compiled by the Spanish Society of Natural Language Processing (SEPLN), which consists of domain generic tweets written in the varieties of Spanish spoken in Spain, Peru and Costa Rica (Sociedad Española para el Procesamiento del Lenguaje Natural, 2018).

The object of analysis of this research is the sentiment analysis component of the DIIA platform, with the aim of discussing its creation and making a proposal for its improvement to reach the baseline levels achieved by state-of-the-art techniques for the sentiment classification task in Spanish (Martínez-Cámara, Martín-Valdivia, Ureña-López & Mitkov, 2015). The hypothesis that underlies the proposed approach is that by adapting the generic classifier to specifically handle Uruguayan Spanish (Rioplatense), and thus developing a localized method, the implementation of the classifier would offer more

accurate and meaningful information about the learning experience of students in the Uruguayan educational context. This approach takes on account linguistic variation, an intrinsic characteristic of all languages that refers to the systematic differences in pronunciation, vocabulary, and grammar of different social and regional groups of speakers of a language (Holmes, 2012). Linguistic variation is a relevant phenomenon for any natural language processing task, and in the case of sentiment analysis it should be considered not only because of the distinctive lexical and syntactic features of the dialect but also because these patterns carry social meanings (Wardhaugh, 2015). Therefore, the sentiment classifier localization process involves mainly enriching it with regional Rioplatense vocabulary and expressions.

The rest of this document is organized as follows: the second section provides the conceptual theoretical framework that supports this study, including the review of the literature on social learning analytics, the sentiment analysis task and its convergence with educational purposes. The third section presents the DIIA architecture and the sentiment analysis component methodology, describing its training, testing, and outcomes. The discussion, implications, results and improvements are provided in the fourth section, focusing on linguistic variation and Uruguayan Spanish. Finally, the last section presents the conclusions and perspectives for future research and development.

2. Related Work

Given the interdisciplinary nature of this study, this section reviews the most relevant related topics to define the scope of the research. First, learning management systems and their uses are described to understand the learning context of the study and to highlight their potential

for applying social learning analytics. In turn, the concepts of learning analytics and social learning analytics are discussed and presented as the learning theories and approaches that support the research. On the other hand, the concepts of artificial intelligence, machine learning, and natural language processing are introduced in order to address the sentiment analysis research area and hence describe the sentiment classification task.

Learning Management Systems

The establishment of the Plan Ceibal in Uruguay has risen interest in leveraging different technologies for the improvement of students' learning experience. As a specific strategy, Learning Management Systems (LMS) platforms have been institutionalized as resources to achieve this goal (Ferrero, Rodríguez, Techera & Motz, 2017). LMSs, also known as content management systems (CMSs) or virtual learning environments (VLEs), are online-based educational systems that give access to educational resources of diverse nature, such as multimedia materials and content; exercises, tasks and assessments; tests and questionnaires, and links to external material, among others (Suero Montero & Suhonen, 2014). What is more, LMSs allow their users to communicate and interact with each other through discussion forums, messaging services and email, chat rooms and blogs (Suero Montero & Suhonen, 2014).

The key element of these systems is the possibility to track student interaction in and with the learning environment, gathering large amounts of descriptive data of users' actions. This service does not only consist of tracking log and browse time, but also includes demographic information such as user profiles; of their progress and academic results; and, remarkably, their interaction data (Buckingham Shum & Ferguson, 2012). Renown examples

of these systems include Blackboard[2] and Moodle[3], and under Plan Ceibal, students and teachers of primary and secondary education use the LMS CREA 2, offered by Schoology[4]. This Uruguayan LMS reported a daily activity of 200,000 active users per day (Plan Ceibal, 2017).

As they are designed and institutionalized for educational purposes, LMSs are considered formal educational platforms. Nonetheless, informal educational platforms are an additional resource for teachers to foster spontaneous interaction with and among their students. These educational environments consist of social networks platforms, where students interact intensively in spaces created autonomously or by their teachers' initiative. Innovative Uruguayan teachers concerned with the improvement of their students' academic performance have explored the use informal social networks such as Facebook, besides the formal learning platforms (LMSs), as venues to stimulate the pursuit and construction of knowledge through interaction between students, as well as between students and the teacher. In this platform, teachers typically create groups for their courses as an open forum and as a channel for exchange of additional information and materials related to their subjects.

The spaces of social interaction just described offer great potential to analyze the way in which students are participating in the construction of knowledge through communication with their peers and teachers. In consequence, both formal and informal learning environment platforms provide meaningful data that may reveal connections between student behavior and specific learning gains (Johnson, Adams Becker, Cummins, Estrada, Freeman & Hall,

[2] www.blackboard.com
[3] www.moodle.org
[4] www.schoology.com

2016). Hence, they create an advantageous opportunity for social learning analytics, learning analytics based on social learning fostered by these environments.

Social Learning Analytics

The Society for Learning Analytics Research (SoLAR)[5] defines learning analytics as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Clow, Ferguson & Brasher, 2015, par. 2). In other words, learning analytics uses learning data, including in some cases big data, to generate actionable intelligence for educational agents, teachers and learners (Ferguson, 2014), for instance "to build better pedagogies, empower active learning, target at-risk student populations, and assess factors affecting completion and student success" (Johnson et al., 2016, p. 38).

Learning in the current digital era is no longer considered an individual activity; rather, it is described as the process of acquiring and updating knowledge from experiences in a dynamic and constant way that occurs when creating learning networks, which can be social connections or large databases (Siemens, 2005). This learning theory, connectivism, is based on social learning, essentially learning through participation, the collaborative construction of knowledge, and its meaning. Social learning is based on the acknowledgment that learning depends to a great extent on social interactions, and thus can be understood as the set of interaction processes that result in viable actions to create change, as occurs with individuals learning within a social context (Blackmore, 2010). This learning based on

[5] http://solaresearch.org/

networks is clearly observed within the framework online social learning environments, where participants, students, and teachers, share information and cooperate to create knowledge.

Accordingly, when learning analytics specifically focuses on the concepts of interaction and collaboration, based on the stance that learning achievements are not merely individual but are developed, carried forward, and passed on collectively, we are talking about of Social Learning Analytics (SLA) (Buckingham Shum & Ferguson, 2012). SLA surpasses summative measurements of students' performance and seeks to return behaviors and patterns in a collaborative learning environment that impact and that indicate an effective learning process (Buckingham Shum & Ferguson, 2012).

Artificial Intelligence, Machine Learning, and Natural Language Processing

As of today, *Artificial Intelligence* (AI) is as ubiquitous in the fields of research and development as in daily life, given its wide range of applications and topics. This includes the automatization of routine labor (Hamid, Smith, & Barzanji, 2017), speech and image understanding (Erden, Velipasalar, Alkar & Cetin, 2016), detection, diagnosis, characterization and monitoring of diseases in medicine (Hosny, Parmar, Quackenbush, Schwartz & Aerts, 2018), and many other tasks that support basic scientific research (Goodfellow, Bengio & Courville, 2016). In general terms, AI may be understood as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation" (Kaplan & Haenlein, 2019, p. 17).

Accordingly, all the problems we attempt to solve with AI involve real-world knowledge, and therefore to tackle them some *learning* must take place. An AI system is said to "learn" when it acquires knowledge through the extraction of patterns from raw data (Goodfellow, Bengio & Courville, 2016), a process known as *Machine Learning* (ML). More specifically, ML consists of the set of automated methods to identify patterns in raw data with the purpose to make predictions (Murphy, 2012) or "perform other kinds of decision making under uncertainty" (p. 1). In other words, ML is the way AI systems learn, and therefore it is the process that allows the development of AI applications themselves.

Among the many areas of application of AI and ML, we find *Natural Language Processing* (NLP). NLP, also referred to as computational linguistics or human language technology, is the manipulation of natural or human language employing computational methods (Bird, Klein, & Loper, 2009) with the goal to allow computers to carry out useful tasks involving language (Jurafsky & Martin, 2008). Uses of NLP include the development of computer dialogue systems (Chen, Liu, Yin & Tang, 2017), machine translation (Gaspari, Almaghout, & Doherty, 2015), question answering (Stroh & Mathur, 2016), and automatic summarization (Nenkova & McKeown, 2012), among many others.

Sentiment Analysis

Sentiment Analysis (SA) or *opinion mining* is the research area of NLP focused on the computational analysis of people's subjective evaluations about entities and their attributes (Liu, 2012; 2015). In other words, it studies the sentiments, opinions, or attitudes people have towards products or services, organizations, individuals, diverse issues, events, or topics. This area has proved useful for a broad range of applications. One of the most popular

examples in the Internet and websites' domain is the use of sentiment analysis techniques for the creation of opinion-aggregation websites, which mainly include product reviews such as movies and electronics, but which could also regard political issues and contain electoral polls, for instance (Pang & Lee, 2008). Sentiment analysis is hence an invaluable resource for business intelligence, uncovering client and general public opinions about their products and services (Pang & Lee, 2008). Moreover, opinion tracking of this nature allows for trend prediction, such as in sales (Yuan, Xu, Li & Lau, 2018) or in the stock market (Al-Augby, Al-musawi & Mezher, 2018). Likewise, sentiment analysis techniques applied to reviews can help to rank products and merchants (Liu, Bi & Fan, 2017); conversely, reputation management and public relations benefit greatly from sentiment analysis (Kharde & Sonawane, 2016). Other computational applications of sentiment analysis techniques include recommendation systems (Chen, Huang, Bau & Chen, 2012), opinion summarization (Yang, Kim & Lee, 2010), and detection of spam or fake opinions (Peng & Zhong, 2014). Furthermore, sentiment analysis systems can facilitate tasks for other sectors, like health and government intelligence. On the one hand, the extraction and treatment of subjective data supports bio-surveillance or monitoring of populations for adverse health issues, such as substance abuse and addiction recovery, self-medication, seasonal events like influenza or environmental allergies, and disease outbreaks, such as the H1N1 virus (Dredze, 2012). In addition, sentiment analysis has served to predict election results (Ramteke, Shah, Godhia, & Shaikh, 2010), detect hostile or negative communications (Kumar, Ojha, Malmasi, & Zampieri, 2018), and characterize social relations (Groh & Hauffa, 2011).

Sentiment analysis and opinion mining are umbrella terms that cover different related tasks, for example review mining (Kamal, 2015), opinion extraction (Ouertatania, Gasmib, & Latiri, 2018), subjectivity analysis (Montoyo, Martínez-Barco, & Balahur, 2012), emotion analysis (Manzoor Hakak, Mohd, Kirmani & Mohd, 2017), affect analysis (Neviarouskaya, Prendinger & Ishizuka, 2010), among many others. One of the most extensively studied tasks of this field is sentiment classification. (Liu & Zhang, 2012). The sentiment classification task has the goal to classify opinion texts according to their polarity, that is the attitude or sentiment of their author with respect to some topic. (Pang & Lee, 2009). This polarity is defined considering either a three-point or five-point scale in a positive-neutral-negative spectrum (Pang & Lee, 2009). More specifically, this task is also referred to as *document-level sentiment classification*, given that the whole text is considered as a single unit and thus it is assumed that the associated polarity represents the sentiments of a single opinion holder towards a single entity (Liu & Zhang, 2012).

The sentiment analysis problem encompasses diverse natural language processing tasks, such as word sense disambiguation (Seifollahi & Shajari, 2019), negation handling (El-Din, 2017), and coreference resolution (Le Thi, Quan & Phan Thi, 2017), for instance. Although this poses considerable obstacles, sentiment classification systems do not require a deep semantic understanding of the analyzed documents, but the grasp of some aspects, namely, the positive, negative or neutral sentiments expressed and their targets (Liu, 2012). To achieve this goal, different methods and classification features have been proposed in the literature.

Methods. Sentiment classification can be treated as a text classification problem with at least two classes, positive and negative. To tackle the task, several approaches may be adopted:

1. **Machine Learning (ML) approach:** This approach involves an ML method and certain features to build a classifier that can associate texts to the sentiment classes (Giachanou & Crestani, 2016). These methods can be divided mainly into two types:

    a. **Supervised Learning Methods:** Supervised learning is the ML technique that requires labeled data which represents the characteristics of a document. Based on that data, it generates a classification model that describes through a mathematical function the relationship between the characteristics of the document and a class (Martínez Cámara 2015). Any existing supervised learning algorithms can be used to sentiment classification (Hajmohammadi, Ibrahim, & Ali Othman, 2012), and some of the most applied are Naïve Bayes (NB), Support Vector Machines (SVM), Maximum Entropy (MaxEnt), Logistic Regression (LR), and Random Forest (RF), among others (Giachanou & Crestani, 2016).

    b. **Unsupervised Learning Methods:** Contrary to supervised learning, this ML technique does not have labeled data a priori, and thus a mathematical classification model cannot be generated. Unsupervised learning methods study the characteristics of each document with the intention of discovering the possible class to which it belongs, mainly recurring to Principal Component Analysis (PCA), clustering and association rule

learning, according to the linguistic characteristics of the documents (Martínez Cámara 2015).

2. **Lexicon-Based (LB) approach:** This approach employs a manually or automatically annotated list of sentiment (positive and negative) terms, to compare against the documents and then determine its polarity (Kharde & Sonawane, 2016). This approach is further categorized into:

   a. **Dictionary-based methods:** A reduced collection of sentiment "seed" words annotated with their polarity is compiled, and following it is grown by searching the synonyms and antonyms of the terms in larger dictionaries or thesaurus, such as WordNet and SentiWordNet (Medhat, Hassan, & Korashy, 2014).

   b. **Corpus-based methods:** Likewise, this method is based on a list of seed sentiment words, however, it is grown by looking for related words in a vast corpus, a collection of texts stored digitally (Lindquist, 2009), according to syntactic patterns (Medhat et al., 2014)

3. **Hybrid *(Machine Learning & Lexicon-Based)* approach:** This comprehensive approach encompasses techniques that combine ML and LB methods (Giachanou & Crestani, 2016).

Features. In the literature, probably a myriad of features has been proposed for sentiment classification, depending on the context of the problem, including the language (or languages) treated, the domain of the texts to classify, and the ultimate purpose of the

classification. Some comprehensive examples of feature categories commonly used in the sentiment analysis task are:

1. **Terms presence and frequency:** Individual words or sequences of N words (called *n-grams*) and their frequency counts. These features may be employed by means of binary weighting (giving a value of one if the word appears and of zero if it does not), or term frequency weighting (Medhat et al., 2014). These weights represent the relative importance of features (Mejova, Srinivasan, 2011).
2. **Part-Of-Speech:** Parts of speech (POS) are the grammatical or syntactic categories of words (Jurafsky & Martin, 2018), such as adjectives, adverbs, and nouns. These POS and some verbs are particularly good indicators of subjectivity and sentiment (Kharde & Sonawane, 2016).
3. **Opinion or sentiment words:** These are words and phrases that convey positive or negative emotions. For example, adjectives like *amazing* and *boring*, adverbs such as *cheerfully* and *slowly*, nouns like *best* and *worst*, or verbs such as *love* and *hate* (Hajmohammadi, Ibrahim, & Ali Othman, 2012).
4. **Negation:** As valence shifters, negative words could invert the opinion (Pang & Lee, 2008), for example, "I like dogs" has a contrary polarity in comparison to "I don't like dogs". However, "not all appearances of explicit negation terms reverse the polarity of the enclosing sentence" (p. 23), as in the example "No wonder this is considered one of the best" (p. 23).
5. **Syntactic dependency:** This feature consists of the order of and relations among words in phrases. Word dependency-based features are generated from

dependency trees or *parsing*, which involves, for instance, the extraction of POS tags (Tubishat, Idris, Abushariah, 2018).

Nevertheless, this is not an exhaustive review of all the methods and features used to tackle the sentiment analysis problem; many other proposals have been put forward in the literature, for instance, the ones presented below in Table 1:

Study	Year	Approach	Method	Features
Khuc, Shivade, Ramnath & Ramanathan	2012	Hybrid	lexicon-based, Online Logistic Regression	Sentiment lexicon, POS, bigrams
Khan, Bashir & Qamar	2014	Hybrid	EEC, IPC, SWNC	Emoticons, positive and negative words, SentiWordNet dictionary
Thelwall, Buckley & Paltoglou	2012	Lexicon-Based	SentiStrength	Emoticons, negations, emphatic lengthening, boosting words etc.
Ortega, Fonseca & Montoyo	2013	Lexicon-Based	clustering-based word sense disambiguation (WSD), lexicon-based classifier	WordNet, SentiWordNet
Saif, He, Fernandez & Alani	2016	Lexicon-Based	SentiCircles	SentiWordNet, MPQA, Thelwall-Lexicon
Ye, Zhang, & Law	2009	Supervised ML	SVM, Naive Bayes, character-based N-gram model	Unigram Frequency
Zhang, Ye, Zhang & Li	2011	Supervised ML	SVM, Naive Bayes	Unigram, Bigrams, Trigrams
Mohammad, Kiritchenko & Zhu	2013	Supervised ML	SVM	Word/character n-grams, POS, caps, lexicons, punctuation, negation, tweet-based
Hamdan, Bechet & Bellot	2013	Supervised ML	SVM, NB	Unigrams, concepts (DBPedia), verb groups/adjectives (WordNet) and senti-features (SentiWordNet)
Dubiau & Ale	2013	Supervised ML	Naïve Bayes, MaxEnt, SVM, Decision Trees, adaptation of Turney's algorithm	presence and frequency of unigrams and bigrams and presence of adjectives
Kiritchenko, Zhu & Mohammad	2014	Supervised ML	linear kernel SVM, MaxEnt	Word/character n-grams, POS, caps, punctuation, emoticons, automatic sentiment lexicons, polarity, emphatic lengthening
Meo & Sulis	2017	Supervised ML	NB, SVM, Random Forest, Logistic Regression	Emotion lexicon, polarity lexicon, latent factor, 5 dictionaries
Prabowo & Thelwall	2009	Supervised ML and Rule-based classification	SVM , Rulebased Classifier	POS tag, Ngrams
Taboada, Brooke, Tofiloski, Voll & Stede	2011	Unsupervised ML	Dictionary based approach	Adjectives , Nouns, verbs , Adverbs, Intensifier , Negation
Dong, Wei, Tan, Tang, Zhou & Xu	2014	Unsupervised ML	AdaRNN	Dependency tree, unigrams, bigrams

Table 1. Approaches, methods, and features for sentiment analysis.

Sentiment Analysis in Educational Research

As it was previously discussed, in today's educational contexts, learning can no longer be studied as individual cognitive or behavioral development; rather, the focus must be shifted towards collaborative processes of knowledge construction (Buckingham Shum & Ferguson, 2012) and the heterogeneous and complex online environments where collaborative learning takes place and the resources required for it (Motz, 2018).

To this end, social learning analytics provides new methods to explore educational data and gain insight into the construction of knowledge through interaction; allowing to get a better understanding of cooperative learning and the relation between student social behavior and specific learning gains. Along the same lines, besides being a crucial element in interaction, it has been shown that language is one of the main tools for learning construction, used by students in accordance with their context, goals, emotions and interpersonal relationships (Wells & Claxton, 2002). The use of language is crucial for knowledge negotiation and construction (Buckingham Shum & Ferguson, 2012), as can be seen in the language used by students in personal communication with other classmates and teachers, because "[t]he ways in which learners engage in dialogue are indicators of how they engage with other learners' ideas, how they compare those ideas with their personal understanding, and how they account for their own point of view" (p. 13). In addition, students use language spontaneously in LMS and social media platforms to express themselves; linguistic productions that carry great knowledge regarding their learning experiences in and outside the classroom whose understanding can "inform institutional decision-making on interventions for at-risk students, improvement of education quality, and

thus enhance student recruitment, retention, and success" (Long & Siemens, 2011, in Chen, Vorvoreanu & Madhavan, 2014, p. 246).

On this account, it has been proven that the analysis of language within educational contexts can provide revealing insight about the learning process and learning experiences of students, and thus is of interest to perform it along with other social learning analytics approaches. In this sense, sentiment analysis techniques such as sentiment classification presents a perfect fit to fulfil this purpose, resulting in its increasing use as a tool for monitoring online learning environments (Harris, Zheng, Kumar & Kinshuk, 2014), Massive Open Online Courses (MOOCs) (Wen, Yang, & Rosé, 2014a; 2014b), social media platforms (Chen, Vorvoreanu & Madhavan, 2014) and online discussion fora (Kagklis, Karatrantou, Tantoula, Panagiotakopoulos & Verykios, 2015).

For example, Kagklis et al. (2015) applied text mining, social network analysis and sentiment analysis techniques to postgraduate students' data from their participation in an online forum. This way, they extracted information about the structure, content of the students' messages and the patterns of interaction among them, but also detected the trend of the sentiment polarity during the course and the progressive students' performance. Hence, students' attitude towards the course in relation to their overall performance was modeled, with the goal to inform tutors and improve the educational process. From this study, it was observed that participation in the forum did not have a significant impact in students' final performance in comparison to the exchanged messages' polarity, which found to have a marginal impact on students' performance.

After a qualitative analysis of students' tweets related to their college life, Chen et al. (2014) identified different problems regarding their educational experiences, mainly derived from heavy study loads, such as the balance between study and life, sleep deprivation, and lack of social engagement. With this data, a multi-label classification algorithm was designed to classify new texts reflecting students' problems, resulting in a trained detector that can be implemented as a monitoring mechanism to detect cases of students at-risk. Ultimately, this application would be to support the decision-making processes of educational administrators and practitioners by conveying insights from the students' learning experiences.

Sentiment analysis has also been used to study the motivation, engagement and dropout risk of students in virtual learning environments, such as in Wen et al. (2014a; 2014b). In their study, they implement an automated fine-grained sentiment analysis in three MOOCs to examine students' opinions trends about their courses and their tools. This analysis allows the study of student motivation and cognitive engagement from the text of forum posts, which the authors correlate with drop out behavior, showing that the more motivation and the more personal interpretation the student expresses, the lower the risk of dropout. Therefore, this kind of sentiment analysis application can serve to detect struggling students and hence provide adequate support.

Along the same lines, Harris et al. (2014) created and implemented a multi-dimensional sentiment analysis agent for LMSs, trained with students' texts from discussion fora, that would provide overall student feedback, and, moreover, identify and alert administrators about striking variations of students' sentiments. In order to do so, the SA agent monitors students' interaction in the LMS and classifies textual data into six categories:

positive, negative, neutral, insightful, angry, and joke. Finally, the authors remark that SA agents of this nature may be particularly useful in larger virtual learning environments such as MOOCs to efficiently inform instructors and administrators about important sentiment changes, and hence properly tackle potential student issues.

3. The DIIA Proposal

The main goal of the DIIA project was to develop a software platform that allowed for the detection of semantic patterns that impact learning, based on students' interactions produced in the online learning environments associated to their courses. This goal was achieved by the DIIA team through the design and implementation of several analytics and visualization services. These services offer teachers and educational administrators different integrated functionalities based on data obtained from the LMSs and informal learning platforms (such as Facebook) with which they already work. In this section, the general architecture of the DIIA platform is first discussed. Then, special focus is given to the visualization aspect of the platform, described in relation to the sentiment analysis module.

General Architecture

The DIIA platform is aimed at supporting teachers' and educational administrator's decision-making by providing insights about students' academic performance and social interaction, as well as the knowledge gained from their participation in educational platforms. The platform was developed following the architectural software pattern Model-View-Controller (MVC) to facilitate the development and maintenance of the components while fostering scalability. The general architecture of DIIA is outlined below in Figure 1.

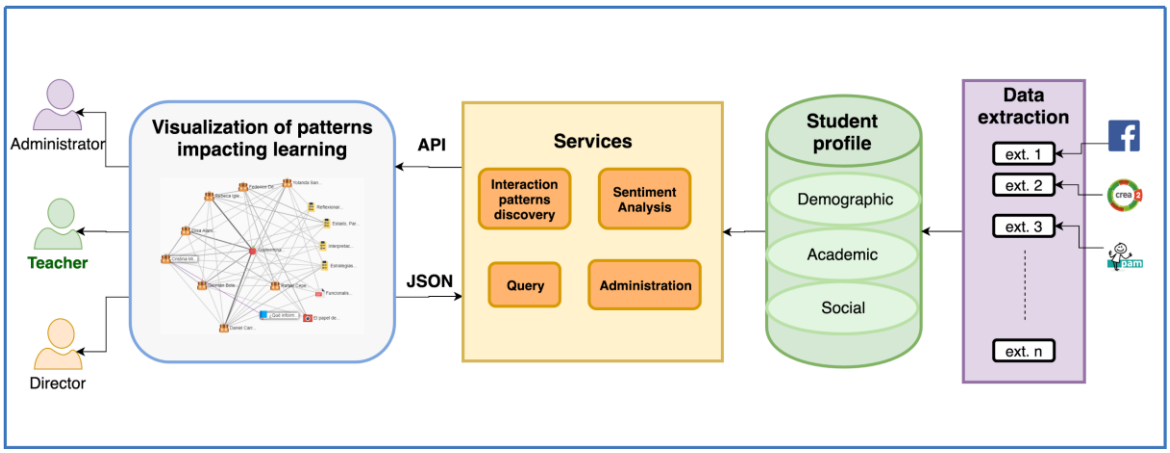

Figure 1. The general architecture of the DIIA platform.

The platform is fed with data derived from formal and informal educational platforms, such as PAM, CREA and the social network platform Facebook. Based on this data, student profiles are modeled, which include demographic, academic, and social interaction aspects. The data extraction component, making use of specialized modules for each type of data source, undertakes the extraction, cleaning and loading the data into the platform database. The data stored in the large DIIA database includes the students' interaction of social nature, among students and teachers, as well as the interactions of students and educational resources and tasks proposed by the teachers. Moreover, the data is sorted according to school cycle, providing historical information suitable for applying pattern discovery techniques.

At the center of the architecture lies the main module that implements the application logic, which provides the services used by the visualization component of the interface. All these services can be easily called by any software technology that supports the HTTP protocol. The transferred data is coded using the JSON format, extensively used in HTTP responses, fostering data reduction over the network and aiding front-end integration. The services the DIIA platform offers include: interaction patterns discovery implementing social

metrics, data query and administration, including the creation, deletion, and manipulation of the principal entities from the database; and sentiment classification.

Platform Visualization

The goal of the DIIA project was to create an analytics and visualization software environment which displayed patterns that impact learning extracted efficiently from different sources, allowing to draw inferences about interaction patterns that impact learning. For its design, Gothelf's (2016) lean UX methodology was followed, which shifts the design focus from deliverables to the actual user experience, achieved only through an iterative process of building and testing minimum viable products and learning from user feedback. The platform was programmed in JavaScript along with React and can be accessed through the project's website. It unifies strategic information of the students' academic performance, interactions, social metrics and sentiment of their texts in an efficient, modern and simple interface.

At the center of the page is the interaction graph, a graph-based representation that permits the detection of patterns from the interactions generated in formal and informal educational platforms (see Figure 2). Hence, the graph is constituted by the subjects and objects of interaction, the teacher, students, resources and activities of the course as the nodes; and the interactions as the edges of different thickness according to the number of interactions. These interactions come from all the LMSs and social network platforms associated with the current course. The nodes are represented with icons: students with backpacks, the teacher with an apple, activities with a clipboard, and the resources are

depicted with different icons depending on their type (for example videos, images, documents).

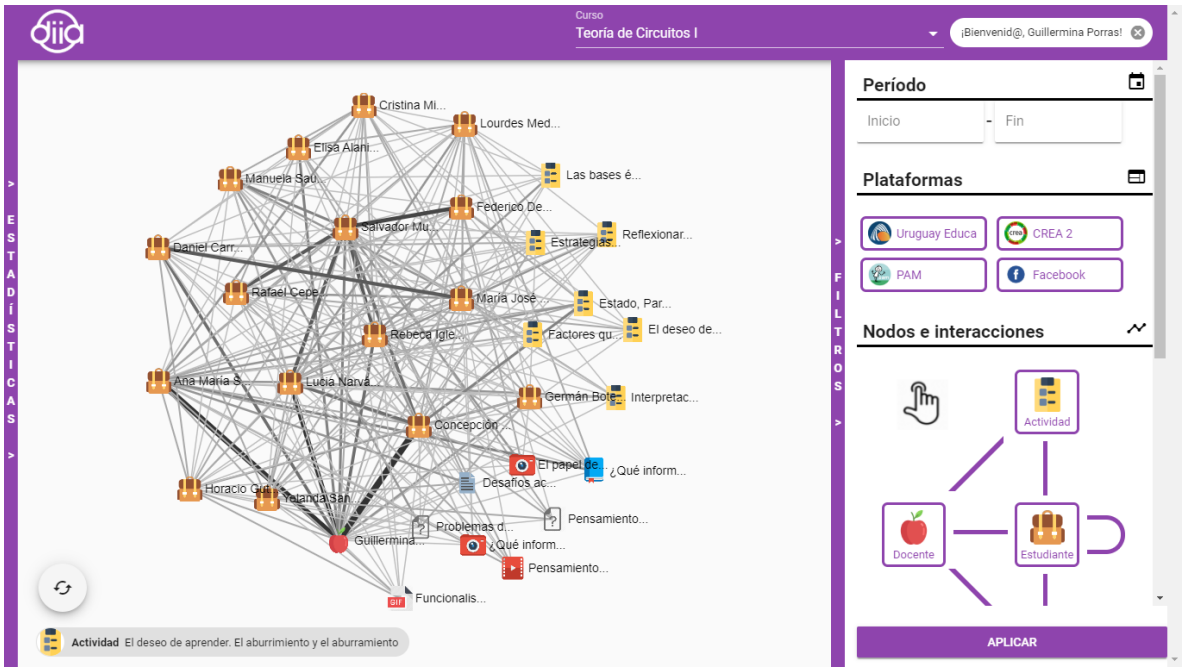

Figure 2. The DIIA platform homepage.

The sidebar at the right of the page (“Filters”) provides filters to visualize different pieces of information in unique ways. It is possible to select the period of time for the information to display, the sources of the information (the educational platforms), which nodes and interactions to display, the types of interactions, the social metrics to apply, and the sentiment of the textual interactions, namely the comments, messages, posts between students or between the students and the teacher. The sentiment analysis filter colors the edges depending on the polarity of the texts, positive (green), negative (red), or neutral (yellow), as shown in Figure 3 below. Furthermore, these interactions can be examined in detail by providing the texts of each interaction and highlighting them in the color of their polarity.

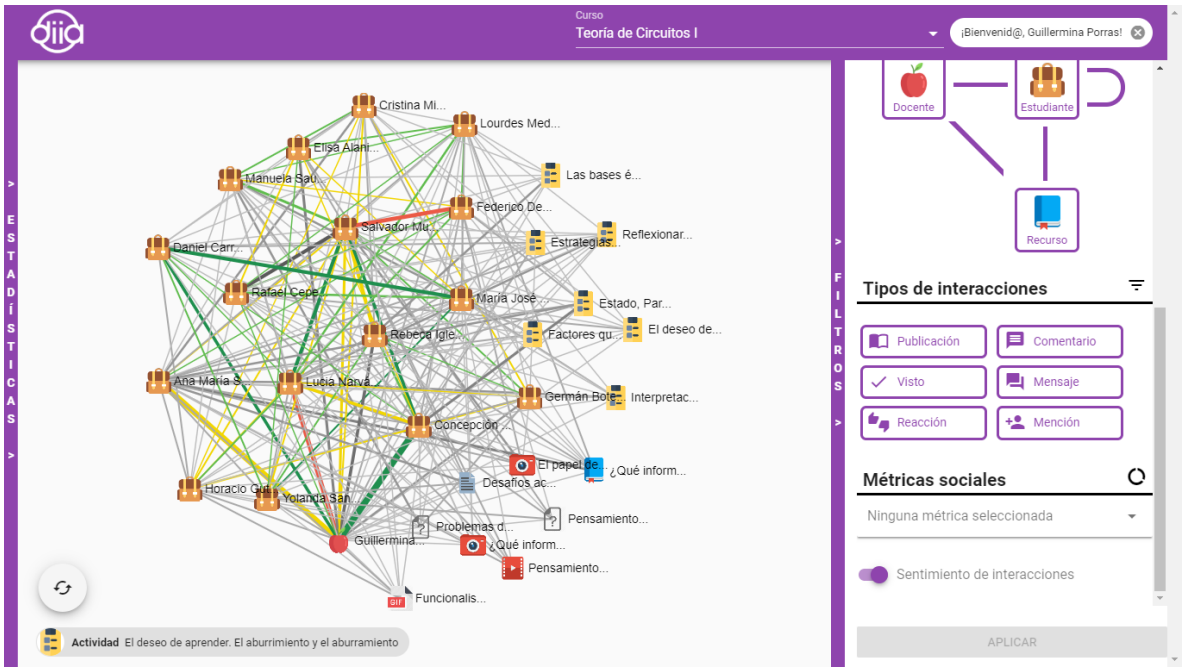

Figure 3. Interactions graph with sentiment polarity filter.

It is essential for teachers to visualize the patterns of interaction that impact the development of the learning community that makes up their courses, which includes the dynamic of the social interactions among their students. The coloring and thickness of the edges according to the polarity and amount of interactions provides a simple and intuitive way to convey this knowledge. This visualization component was designed to support the interpretation of teachers about the motivation, engagement, and relationships among students, and furthermore, it helps them to become aware of possible risk situations such as bullying, low self-esteem, isolation, and sexual harassment that may be found in students' online interaction. However, this component also allows finding possible opportunities to stimulate students' interest, foster teamwork, and many other actions to improve the learning experience of their students.

4. The DIIA Sentiment Analysis Methodology

The DIIA project had the goal to create a software service to discover semantic patterns impacting learning based on students' interactions in social networks. A great part of the interactions involves text; documents such as posts, comments, and messages generated through the use of formal and informal learning platforms by students and teachers. Therefore, among the different tools created to analyze the interactions associated to the elements of the educational networks (teacher, students, resources and tasks), a sentiment analysis module was created by the DIIA team, using the Python programming language[6]. This classification component processes the subjective textual information generated from social interactions effectively and performs semantic analysis to predict the negative, positive or neutral polarity of the documents. The sentiment classification model was created under a supervised learning approach and classifies the texts based on their lexical-syntactic structure, using the 150 most frequent word trigrams in a vector space representation of their frequency of occurrence represented as vectors of three-words windows (called trigrams) as the classification features; and uses the classical classification algorithm Support Vector Machine (SVM).

In this section, the key elements associated to the sentiment classifier methodology are described, emphasizing the dataset selection, text preprocessing, feature selection, the

[6] www.python.org

training and evaluation processes of the model, and the results achieved. In addition, the full implementation of the module can be found in the Github repository of the project[7].

Dataset Selection

For the training of the sentiment classifier, the DIIA team used the InterTASS-2017 corpus compiled by the Spanish Society of Natural Language Processing (SEPLN), considering their expertise associated to the creation and annotation of Spanish language datasets, the free availability of the data, and the origin and the inter-varietal nature of the selected dataset. This corpus is composed by tweets written in the Spanish varieties from Spain, Peru and Costa Rica (Sociedad Española para el Procesamiento del Lenguaje Natural, 2018), in contrast with most of the corpora available that consist mainly on Castilian Spanish. Given that the classifier is thought for its application to the Uruguayan Spanish regional dialect, to have such a varied corpus provides a diversity of lexical-syntactic structures that might be present in the input texts. Furthermore, the texts from the InterTASS-2017 corpus were generated from the social network platform Twitter, which presupposes a more spontaneous language, similar to the messages, posts, and comments shared by students in the formal and informal educational platforms.

The InterTASS-2017 corpus is originally in XML format[8] and is annotated with four sentiments/polarities: Positive (P), Negative (N), Neutral (NEU), and none of the above (NONE). It is divided into a training, development and test sets which consist of 1008, 506 and 1899 tweets respectively. However, for the creation of the sentiment classifier, it was

[7] https://github.com/GrupoDIIA/Sentiment-Analysis-for-DIIA
[8] https://www.w3.org/TR/xml/#sec-intro

decided to use only the training and test sets for the training and testing of the module. Additionally, the tweets without a polarity (those annotated as "NONE") were discarded because they could introduce noise in the basic classification process, given the possible similarities these documents could have with the ones annotated as neutral. The main properties of the dataset are shown in Table 2 below.

Features	***Training set***	***Test set***
Dataset main source	Twitter	
Number of texts	869	1625
Positive texts (P)	318	642
Negative texts (N)	418	767
Neutral texts (NEU)	133	216
Average words per text	68.7	72.8
Vocabulary size	10456	16745

Table 2. Main properties of the InterTASS-2017.

Dataset Preprocessing

Before the creation of the model, it is necessary to adapt, clean and eliminate redundant or noisy information from the dataset texts. The InterTASS-2017 corpus was preprocessed by performing the following actions:

1. The dataset documents were converted from their original XML format into a plain text format for its handling.
2. The data was cleaned: punctuation, diacritical marks and all the elements that are not part of the ASCII encoding were removed. Additionally, the elements associated with Twitter texts were removed, namely URLs, hashtags (#) and user mentions (@).
3. Afterwards, all the words were changed to lowercase.

4. Finally, as mentioned previously, all the documents with the "NONE" polarity tag were removed from the dataset in order to have representative examples of the three main sentiments (P, N, and NEU) without introducing noise into the classification model.

Feature Selection and Representation

After the dataset is preprocessed and ready for handling, the next step is the extraction of representative features from texts, for the construction of a classification model using a machine learning technique to predict the polarity of documents.

As discussed in Section 2, there are probably infinite linguistic elements that can be extracted from texts and, hence, that can be used to represent them. In the case of the DIIA sentiment classifier, the development team decided to use word n-grams, given that different studies related to the sentiment analysis have shown that this kind of textual features helps to capture the writing style of documents (Aisopos, Papadakis, Varvarigou, 2011; Deng, Sinha, Zhao, 2016). Accordingly, the documents were represented with word trigrams in a vector space representation (Hladka & Holub, 2015), where the frequency of occurrence of the most recurring elements for each document in the dataset is quantified.

DIIA's Sentiment Classifier Using a Supervised Learning Approach

Once the linguistic features to be extracted and the type of representation have been decided, a supervised learning approach can be used for the construction of the model classifier. This type of learning technique contemplates two major stages: a training and test phase (Harrington, 2012), as shown in Figure 4.

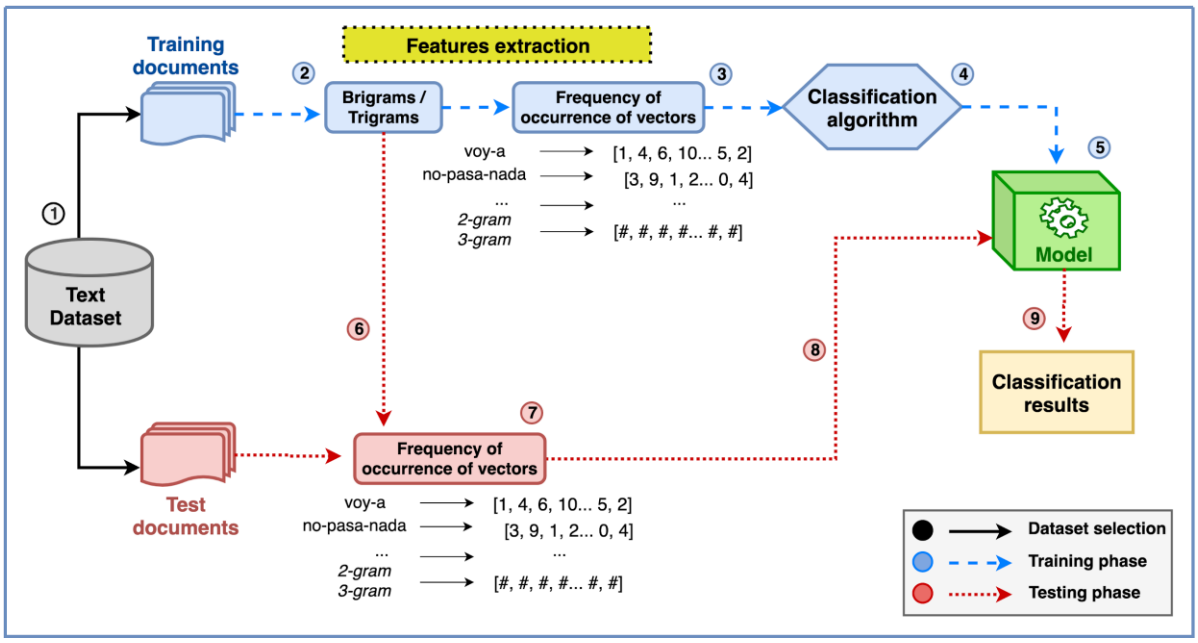

Figure 4. Sentiment classification using a supervised learning approach.

In the training phase, the corpus' documents labeled with their polarity (1) are used in the form of trigram vectors (2,3) to train a classification algorithm (4), in order to create a model (5) that can predict the sentiment (positive, negative or neutral) associated to a text. Next, in the test phase, the input documents are not annotated with their polarity. These may be documents that are not part of the dataset (unseen text samples). The texts are then transformed into their trigram vector representation (6,7) and are given to the previously built model (8). The result of the testing is the corresponding polarity label (9) associated with each document, which can be used to evaluate the performance of the model.

On the same line, to build the model, the classification algorithm Support Vector Machine (SVM) was used. The DIIA team selected the SVM algotithm considering the strong performance obtained with its implementation in several sentiment analysis tasks (Medhat et al., 2014).

Evaluation and Results

Finally, the DIIA model was evaluated against the test dataset partition, an evaluation metric to assess the effectiveness of the predicted classification results. Since it is the most frequently used, the accuracy evaluation metric was chosen, which measures the percentage of correct predictions of the model (Giachanou & Crestani, 2016). In accordance, the DIIA sentiment classifier obtained a model accuracy of 0.472, a considerably lower figure in comparison with the baseline levels reached by similar models and other state-of-the-art techniques for the sentiment analysis problem (Hussein, 2016; Giachanou & Crestani, 2016) and for the sentiment classification task in Spanish (Martínez-Cámara, Martín-Valdivia, Ureña-López & Mitkov, 2015). Therefore, these results create the opportunity to revisit the model and propose improvements, which is the objective of the present study and which will be the focus of the following section.

5. Linguistic Framework for the Localization Proposal

Aligned with the goal of shedding some light on the semantic patterns that impact learning within the framework of the DIIA project, a sentiment classifier was developed by the team. This system would determine the positive, negative or neutral sentiment of texts written by students in formal and informal educational platforms, based on their lexical-syntactic structure and using a supervised learning technique. However, the results obtained by the method did not meet the baseline precision levels for this task, and it is on that account that the present study outlines a proposal for improvement from a linguistic perspective: to adapt the generic classifier to handle Uruguayan (Rioplatense) Spanish specifically and thus develop a localized method. This way, the implementation of the classifier would offer

accurate and meaningful insights into students' learning experience in the Uruguayan educational context. Specifically, I implemented four localized model approaches, trained in a dialect-specific dataset and involving several text representation features and different machine learning algorithms. The reasoning behind my proposal, theoretically founded in the linguistic variation phenomenon, is presented in this section.

Linguistic Variation

Linguistic variation is an intrinsic characteristic of all languages which refers to the systematic differences in pronunciation, vocabulary, and grammar of different social and regional groups of speakers of a language (Holmes, 2012). Each of these groups speaks a *dialect* of that language, "mutually intelligible forms of a language that differ in systematic ways" (Fromkin, Rodman & Hyams, 2011, p. 430). It is important to highlight that, therefore, a language is a collection of dialects, and thus a dialect is not an inferior, simpler or corrupt form of a language nor is any variety linguistically superior to any other (Fromkin et al., 2011; Holmes, 2012; Wardhaugh, 2015). Further, the linguistic features carried by dialects convey social meanings and distinguish the groups from one another (Wardhaugh, 2015).

Linguistic variation develops when some physical or social communication barrier separates groups of speakers and hence the changes to linguistic properties of their language do not spread across them, resulting in the rise of more profound differences between them. As a result, dialects emerge. In particular, when several linguistic distinctions concentrate in a specific geographic region, the language involved becomes a regional dialect (Fromkin et al., 2011). This is the case of Uruguayan or *Rioplatense* Spanish, whose linguistic particularities are discussed as follows.

Spanish in Uruguay

According to the Instituto Cervantes (2018), Spanish is the second most spoken language in the world by number of native speakers, and also the second language of international communication, with a total number of speakers that surpasses the 577 million worldwide, making up almost 8% of the world's population. It is the official language of 21 countries (Instituto Cervantes, 2018), which entails great regional variation across the globe.

Uruguay is a small South American country with a population of nearly 3.5 million people, of which 98.4% speaks Spanish as their native language (Instituto Cervantes, 2018). Uruguay shares borders with Brazil to the north and east, and with Argentina to the west, separated in the south by the Río de la Plata (River of Silver). It is in the area of the river's basin, a large part of Argentina and in the whole of Uruguay, that the *Rioplatense Spanish* dialect is spoken.

Rioplatense Spanish differs from most Spanish variants mainly because of the history of conquest of Uruguay and the consequential influence of other languages. Uruguay lived a late and discontinuous colonization and a brief colonial period, which, together with its geography, contributed to make it a region slightly isolated from the peninsular cultural heritage (Bertolotti & Coll, 2006). Upon the arrival of the European settlers, there was close contact and linguistic interaction with the indigenous inhabitants, resulting in particular vocabulary terms of current Rioplatense Spanish, namely Guaraní terms for toponymy, fauna and flora; and many everyday language terms from Quechua origin (Bertolotti & Coll, 2006). However, given the generalized Hispanization process and extermination of the original groups, today no indigenous languages are spoken in the country, in contrast to the rest of

Hispanic America where important sectors of the population still preserve their native languages (Bertolotti & Coll, 2006).

On the other hand, other European languages made their way into Uruguay and had a significant linguistic influence. During the colony, Portuguese entered the country along with Spanish (Elizaincín, 2009) and still coexists with Spanish in the north of Uruguay in a *diglossic* situation, which means that both languages exist side-by-side in the community but are used in a complementary way for different functions and in different domains (Wei, 2012). In this sense, Spanish is the language of education, administration and of most services, while Portuguese is used at home and in more familiar registers (Behares 2007). Nevertheless, this Portuguese is not the same as in Portugal nor in Brazil, rather, it has become "border Portuguese" or *Portuñol*, characterized by rural Brazilian Portuguese features, Spanish interferences, and hybrid forms of Spanish and Portuguese (Carvalho, 2003).

On the other hand, the Italian language was brought into Uruguay by migratory waves during the 19th century, which spread from Montevideo in the South inside the country, namely into the center and the west (Palacios, 2015). Italian had a great linguistic influence not only because of the incorporation of lexicon and its effects on intonational patterns, but also on some morphological and syntactic aspects, such as in the verbal paradigm (Bertolotti & Coll, 2014).

Several linguistic features distinguish Rioplatense Spanish from other Spanish dialects. Some of the most noticeable differences are phonetic, such as *seseo* and *yeísmo*, which both involve the loss of distinction between phonemes; between the voiceless

interdental fricative /θ/ and voiceless alveolar fricative /s/, and between the lateral /ʎ/ and palatal /y/ phonemes, respectively (Bertolotti & Coll, 2014). However, there are other grammatical attributes that characterize this variant and that are relevant for the present study, since they are manifested in the written discourse.

Perhaps the most distinctive feature of Rioplatense Spanish is *voseo*, the use of pronominal or (modified) verbal forms of the second person of the plural *vos* to address a single interlocutor, whose use denotes closeness and familiarity (Real Academia Española, 2009). The first case, pronominal *voseo*, involves the use of *vos* as the pronoun of the second person singular instead of *tú* and *ti*. This means that *vos* is used as a subject, as a vocative, with a preposition, and as an object of comparison. However, for the clitic and possessive pronouns the forms of *tuteo* (use of the *tú* pronoun for the second person singular) *te*, *tu*, and *tuyo* are used (Real Academia Española, 2009). On the other hand, "verbal *voseo*" consists of the use of modified verbal endings or suffixes proper to the second person plural *vosotros*, for the conjugated forms of the second person singular, regardless of the pronoun used: *vos/tú comés, vos/tú comís* (you eat, you eat) (Real Academia Española, 2009). These modified conjugations are different for each tense but also vary according to social and regional factors. In Rioplatense Spanish, the verbal paradigm is constituted by *vos* forms with reduction of the diphthong in the indicative present (*cantás, comés, vivís*), by the *vos* forms of the imperative (*cantá, comé, viví*) and by the forms of *tuteo* for the rest of the verb tenses (Bertolotti & Coll, 2014). Particularly, Uruguayan Spanish characterizes for having three modalities or combinations of pronominal and verbal forms of *tuteo* and *voseo* (Bertolotti & Coll, 2014):

1. **Pronominal *tuteo* and verbal *tuteo*:** the pronoun *tú* is accompanied by verbal forms of *tuteo*. For example, "*Y tú aprendiste de todo, supongo…*" (Dabazies, 2003) (And you learned about everything, I guess...). This modality is used in more formal and respectful registers, and is considered as a more prestigious or correct form.
2. **Pronominal *voseo* and verbal *voseo*:** the pronoun *vos* is accompanied by verbal forms of *voseo*. For example, "*Yo sé que vos aguantás*" (Galeano, 1979) (I know that you can hold on). This is the most extended form, and is used in familiar and intimate contexts.
3. **Pronominal *tuteo* and verbal *voseo*:** the pronoun *tú* is used with the verbal forms of *voseo*. For example, *"No, tú no podés haberte ido con ellos"* (Plaza Noblía, 1991) (No, you couldn't have gone with them.). This hybrid combination signals closeness through the verbal *voseo* and, at the same time, denotes deference through the pronominal *tuteo*.

Although each of these combinations exists in other varieties of Spanish, the combination of the three in the same dialect distinguishes Uruguay in the Spanish-speaking linguistic landscape (Bertolotti & Coll, 2014).

Likewise, another form of treatment typical of Rioplatense Spanish is *che*, a word of Guaraní origin. In this language, the form has pronominal use different to those in Spanish, and although it is formally categorized as an interjection (Real Academia Española, 2019), it mostly serves the function of a singular and plural vocative (Bertolotti & Coll, 2006). In these cases, it is usually used with a noun in apposition, for instance: "*Pero, che, Mariano, creía*

que éramos amigos…" (Chavarría, 2002) (But, *che*, Mariano, I thought we were friends...). It is used to call, stop or ask someone for attention, or to denote amazement or surprise (Real Academia Española, 2019).

Another linguistic feature of Rioplatense Spanish that interests this study is the form and function of diminutives of nouns. This construction is made by the use of the suffix *–ito* and it is principally used to convey lessening or smallness, nevertheless they often express different qualities and degrees of appreciation (Merriam-Webster, 2019). In this sense, the diminutive can be used to signal fondness or affection, as in "*bebito*" (the diminutive of *bebé*, baby), and does not refer to the size of the baby (Bertolotti & Coll, 2014). Conversely, using the diminutive of *marido* (husband), "*maridito*", may be interpreted as implying that that husband "lacks some of the prototypical conditions of a good husband" (Bertolotti & Coll, 2014, p. 33). Furthermore, as Bertolotti & Coll (2014) explain, in Uruguayan Spanish this suffixation may even create a new word, for example "*cochecito*" (the diminutive of *coche*, car) does not refer to a small car, rather, to a stroller.

Similarly, *re-* and *super-* are appreciative prefixes distinctive of Rioplatense Spanish that modify adjectives with the purpose of intensifying their meaning, for example, its use with the adjectives "*reloco*" (very crazy) or "*superlindo*" (super or very cute) (Palacios, 2015). It is especially interesting that these appreciative prefixes are also used colloquially with adverbs and verbs, for example with "*remal*" ("very" bad) or "*lo reamo*" ("I love him a lot"), serving as modalizer quantifiers that allow speakers to "resize reality in a different way" (p. 334).

Lastly, in the lexicon of a dialect is where we best find variety reflected. Spanish in particular has a history of being influenced by the most varied linguistic sources, and further, its distribution across the globe accounts for its great diversity given geographical, cultural and sociological factors. Some of the most defining lexicon of Rioplatense Spanish is of Peninsular origin, words no longer used in Spain or that have had a semantic shift (Palacios, 2015), for example *pollera* (skirt, "*falda*" in Spain), *vereda* (sidewalk, "*acera*"), *frutilla* (strawberry, "*fresa*"). Likewise, words from Italian origin may be found in this variety, such as *nono/a* (grandparent, "*abuelo/a*" in most Spanish dialects), *pibe* (kid, "*muchacho/a*"), or the distinctive greeting *chau* (Palacios, 2015). Finally, Rioplatense Spanish shares lexicon from different indigenous languages origins with other Hispanic American dialects, for example *maní* (peanut, "cacahuete" in Spain), *quirquincho* (armadillo, "*armadillo*"), and *ananá* (pineapple, "*piña*"), among many others (Palacios, 2015).

Now that the phenomenon of linguistic variation has been discussed and the distinctive linguistic features of Uruguayan Spanish of most relevance to the scope of this study have been described, the next section presents my proposal of localizing the sentiment analysis module of the DIIA platform. The goal is that these factors serve as a guideline for the modifications to the model in order to provide real and revealing insights into students' learning experience in the Uruguayan educational context.

6. Sentiment Classifier Localization Methodology

My linguistic localization proposal comprises different strategies for the improvement of the model, and to explore its feasibility I compared the proposed approach with other classical methods applied for solving related text classification tasks. I implemented four localization

approaches exploring several text representation features including n-grams, POS tags, and a variety of stylistic features. By the same token, I used different machine learning algorithms, such as SVM, Naïve Bayes, logistic regression and a decision tree. Nonetheless, the main change for the adaptation of the classifier is training the proposed approaches on a new, dialect-specific dataset that includes Uruguayan Spanish texts. This would enrich the model with regional vocabulary and expressions, and other linguistic characteristics representative of this regional dialect, such as morphosyntactic features. In this section, the key elements associated to the sentiment classifier methodology are described, emphasizing the dataset selection, text preprocessing, feature selection, the training and evaluation processes of the model, and the results achieved.

Dataset Selection

The sentiment classification model built for the learning analytics platform DIIA predicts the positive, negative or neutral polarity of texts based on their lexical-syntactic structure, represented as vectors of trigrams, and uses the classification algorithm SVM. This method follows a supervised learning approach, which entails that the model is based on labeled data representing the characteristics of the documents to classify. The DIIA model was trained on a sort of "international" Spanish corpus, a dataset composed by tweets written in the regional varieties of Spanish from Spain, Peru, and Costa Rica. However, the context of implementation of the sentiment classifier and the DIIA platform is in Uruguay, and although the language of the four countries is Spanish and the corpus includes Latin American varieties, each region has its own dialect with its own different representative linguistic features, as discussed in the previous section. Through this study, I argue that it is from this

initial step that the existing classifier faces a significant challenge because it was trained on a set of documents that do not reflect the linguistic characteristics of Uruguayan Spanish. Therefore, the examples from which the model is learning are not representative of the texts it should classify.

In consequence, the decisive element of my localization proposal is training the model on a Rioplatense Spanish dataset. Ideally, this corpus would be composed of texts generated by Uruguayan students through their interactions with and in formal and informal educational platforms, namely CREA 2 and Facebook. This way, the linguistic characteristics of the Uruguayan dialect would be represented, accounting also for the nature and format of the publications, comments and other messages written in educational platforms and social networks. Due to the fact that the DIIA project is in a prospective phase, access to this type of data could not be granted for the compilation of a training data set. Notwithstanding, within the framework of the present study, I conducted experiments using a Uruguayan corpus created by Mori, Tambucho and Cardozo (2016) to train a sentiment analysis system that would serve to carry out a reputation study based on comments extracted from the social network Twitter.

The corpus consists of 2466 tweets taken from Uruguayan accounts, originally annotated with four sentiments/polarities: Positive (P), Negative (N), Neutral (NEU), and none of the above (NONE). In order to compile it, Mori et al. (2016) chose Uruguayan accounts and diverse controversial topics, popular during the period of corpus preparation, ranging in the domains of politics, sports and international events. Accordingly, these topics elicit positive, negative and neutral opinions from the users. The authors downloaded the

tweets using the Twitter API from the selected accounts and also by searching related hashtags or trending topics and user mentions. Once the tweets were downloaded, multiple users participated in a voting process to classify them through a web and another mobile application designed for this purpose.

Given that this corpus consists of a single dataset (unlike the InterTASS-2017), I performed a K-fold cross-validation technique for the training and testing of the proposed configurations. This procedure consists of randomly splitting the training set into *K* distinct subsets or folds, training and evaluating the model *K* times, using a different fold for evaluation every time and training on the other *K*-1 folds (Géron, 2017).

In addition, given the scarcity of neutral examples, I merged the tweets annotated as "NONE" with the ones of neutral polarity, following the methodology of the authors (Mori et al., 2016). The main properties of the dataset are shown in Table 3 below.

Features	***Dataset***
Dataset main source	Twitter
Number of texts	2466
Positive texts (P)	618
Negative texts (N)	652
Neutral texts (NEU)	1196
Average words per text	15.82
Vocabulary size	9219

Table 3. Main properties of the Uruguayan dataset by Mori, Tambucho and Cardozo (2016)

Dataset Preprocessing

For the localization of the model, the preprocessing of the Uruguayan Spanish dataset should be different than the one of the InterTASS-2017 corpus, performed for the DIIA sentiment classifier. To render the linguistic characteristics of the Uruguayan dialect and the

representative features of the texts generated in educational and social platforms, the corpus must remain almost intact. To avoid encoding problems, I removed diacritical marks, yet preserved other UTF-8 characters, such as *emojis*. Similarly, the elements associated with writing on social and educational networking platforms such as URLs, hashtags and user mentions are maintained but replaced by an identifier: "http", "#" and "@", respectively. By the same token, I maintained the original word casing and punctuation. This way, the stylistic and formatting features of the documents are represented, which have been shown to accurately characterize writing styles (Laboreiro, Sarmento & Oliveira, 2011). Lastly, as mentioned previously, I recategorized all the documents with the "NONE" polarity tag as "NEU" in order to have representative examples of the three main sentiments (P, N, and NEU).

Feature Selection and Representation

In order to adequately localize the sentiment classifier model, for it to handle students' documents from educational platforms and provide insights into the sentiments they express and thus into their learning experiences, the extraction of features that truly represent the texts is crucial. To integrally capture the linguistic features of Uruguayan Spanish and the characteristic elements of the educational and social platforms texts', I propose several features and text representations. These configurations are outlined below, and the corresponding experiments are described afterwards.

Original DIIA feature engineering approach. The first scheme proposed is to replicate the feature extraction scheme and the text representation used for the DIIA sentiment classifier model. These were the 150 most frequent word trigrams in a vector space

representation of their frequency of occurrence. However, in this instance, the n-grams would be built from the Uruguayan dataset. The steps taken to generate this text representation are shown by Figure 5 and described below.

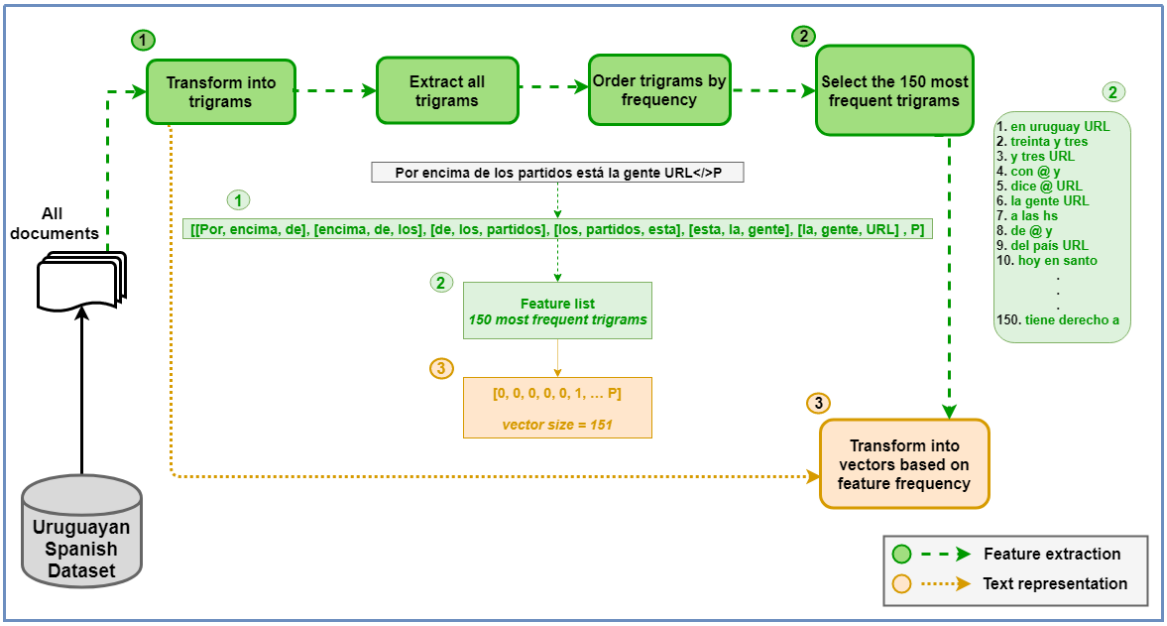

Figure 5. Original DIIA feature engineering approach.

First, each document was transformed into three-word windows called trigrams (1). After all the trigrams of the dataset were extracted, they were ordered by frequency and the 150 most frequent trigrams were identified and selected as features (2). Finally, each document was transformed into a vector space representation based on the frequency of occurrence of the 150 trigrams in the message (3).

Most frequent content words approach. Similarly to the first proposition, it is posited to represent the documents as vectors of some of the most frequent tokens of the collection, but for this configuration these would be the most frequent content words, the ones that carry specific semantic content and hence convey the principal meanings of sentences, such as nouns, verbs and adjectives; as opposed to the function words, the ones that fulfill a merely

grammatical role (Corver & van Riemsdijk, 2013), such as prepositions and articles. This approach is depicted in Figure 6 and described as follows.

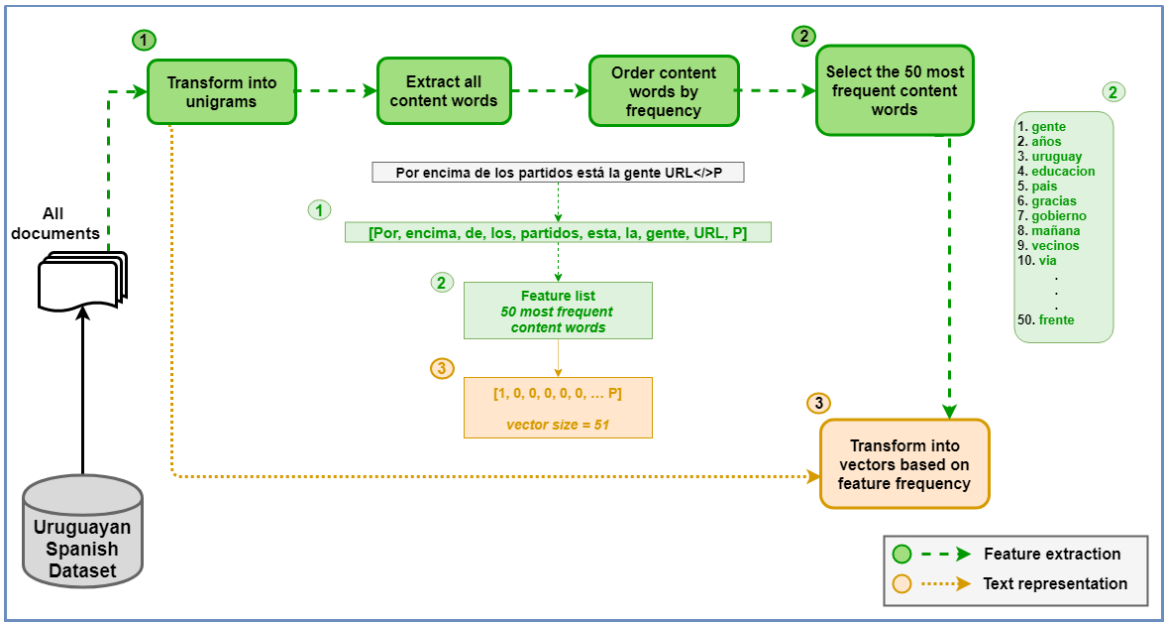

Figure 6. Most frequent content words approach.

First, each document was transformed into one-word windows called unigrams (1); in other words, the documents were divided word-by-word. Then, all the content words of the dataset were extracted and ordered by frequency to obtain the 50 most frequent content words and use them as features (2). Lastly, each document was transformed into a vector space representation based on the frequency of occurrence of the 50 content words in it (3).

Stylistic features approach. As suggested previously, stylistic features may help to profile authors and styles, and therefore it is proposed to use elements of this nature as features in a vector space representation to render the sentiment documents. These include word casing, punctuation, repetition of characters, presence of emoticons and emojis, graphic signs that represent facial expressions through ASCII symbols and Unicode graphic symbols used to express concepts and ideas, respectively (Novak, Smailović, Sluban, & Mozetič,

2015); presence of URLs, and frequency of user mentions and hashtags. The methodology of this text representation approach is displayed in Figure 7 and explained below.

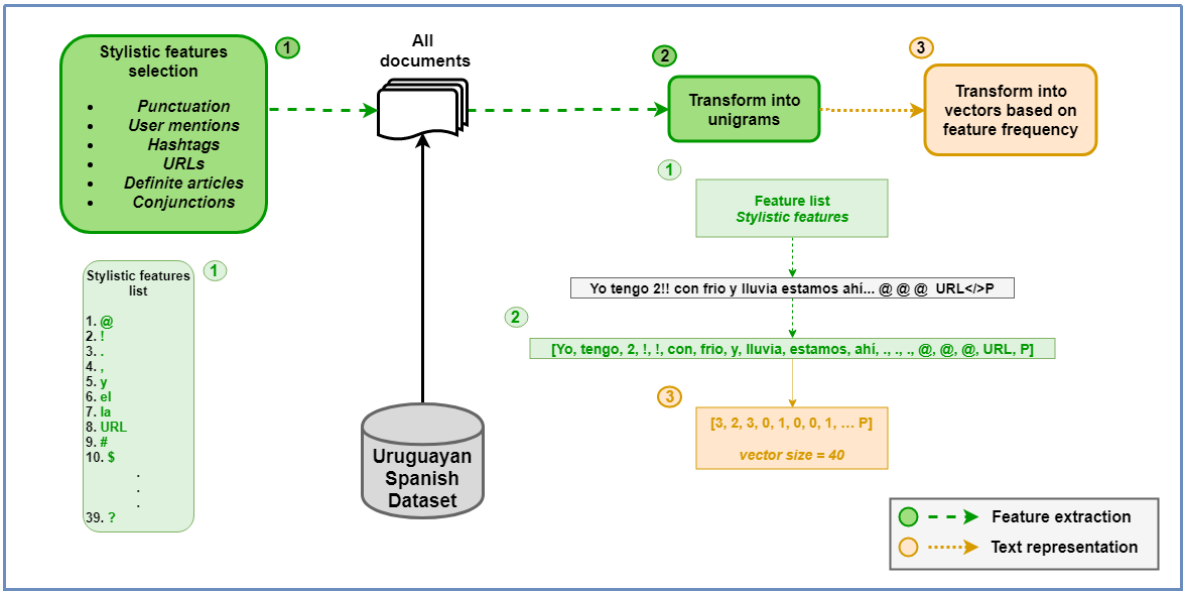

Figure 7. Stylistic features approach.

First, the feature set was defined, including the stylistic elements of punctuation, user mentions, hashtags, URLs, definite articles, and conjunctions; comprising a list of 39 elements (1). Next, each document was transformed into unigrams (2). Afterwards, each document was transformed into a vector space representation based on the frequency of occurrence of the 39 stylistic features in the message (3).

Part-Of-Speech (POS) approach. Finally, it is argued that probably the most efficient way to capture the linguistic particularities of the Uruguayan dialect is to represent the texts with their grammatical categories and syntactic relations. This proposal involves parsing the documents to represent each of them with their POS, and then use several the grammatical categories as features: first-person pronouns, verbs, adjectives, and adverbs. These POS have proved to be reliable indicators of sentiment (Kharde & Sonawane, 2016), and personal

phrases (sentences having a first-person pronoun) have also been shown to integrate the essence of subjective texts (Ortega-Mendoza & López-Monroy, 2018). Moreover, this feature representation scheme would ideally involve parsing meticulously (almost tailor made for Rioplatense Spanish), to analyze in depth the morphosyntactic aspects of the texts. Especially, information about the modalities of *voseo* and *tuteo*, and the use of diminutives and appreciative prefixes would most likely provide revealing insights about the sentiment expressed in the documents. The process behind the implementation of this approach is shown in Figure 8 and described as follows.

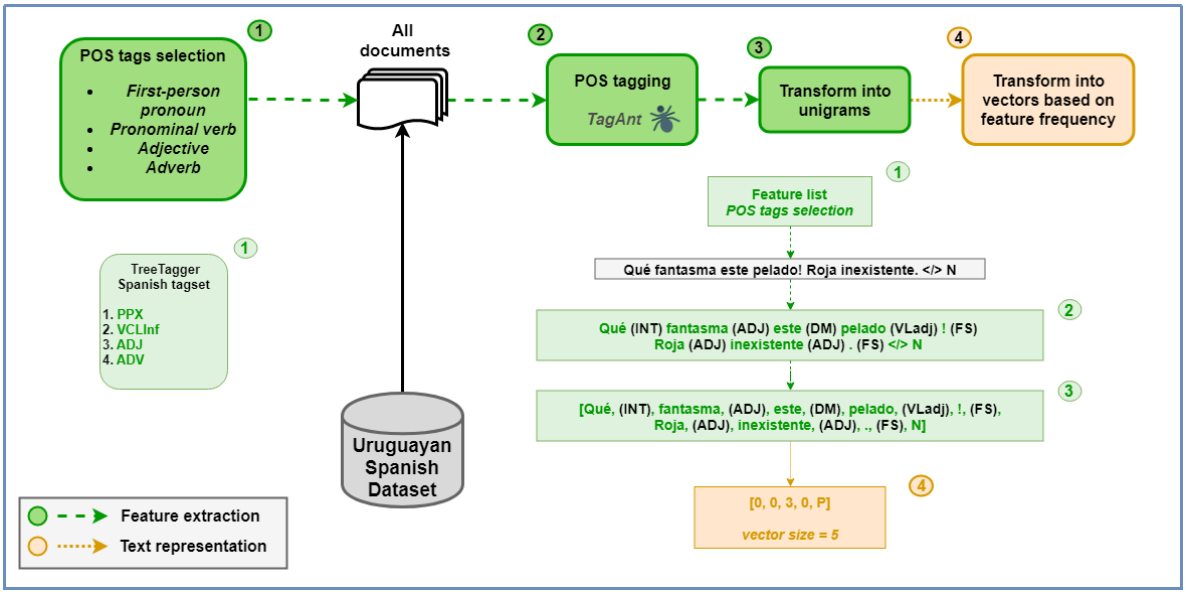

Figure 8. Part-Of-Speech (POS) approach.

First, the list of features was defined, including the grammatical categories related to personal phrases, namely the POS tags for first-person pronouns, verbs, adjectives, and adverbs (1). Second, all documents were parsed; each word was tagged with its corresponding POS (2), using the TagAnt[9] Part-Of-Speech (POS) tagger open software built

[9] http://www.laurenceanthony.net/software/tagant/

on TreeTagger developed by Schmid (1995). The TreeTagger's tagset used was developed for corpus annotation in Spanish, consisting of 75 tags and included in the Appendix. Following, each message was transformed into unigrams (3). Finally, each document was transformed into a vector space representation based on the frequency of occurrence of the four POS tags in the document (4).

Localized Sentiment Classification Model

This section discusses the localized sentiment classification model, adapted to specifically handle Rioplatense Spanish. The classifier is a system capable of processing the subjective textual information generated from social interactions in the Uruguayan educational context and performing semantic analysis to predict the negative, positive or neutral polarity of the documents. The model was developed using the Python programming language, along several software tools, described below.

Similar to the DIIA method, the model classifier was developed using the Python programming language, and a supervised learning approach is proposed for its construction. However, this method incorporates the four varieties of features and representations proposed for the localization. The proposed approaches were tested by means of diverse machine learning algorithms used for solving sentiment analysis tasks. Moreover, as it was previously stated, a K-fold cross-validation technique was performed for the training and testing of the proposed configurations. For the experiments, this technique employed 10-folds. These procedures and the classification itself were done by means of the "Weka" data mining open

software[10]. The design of the prototypical model for the different experiment configurations is shown in Figure 9 and described below.

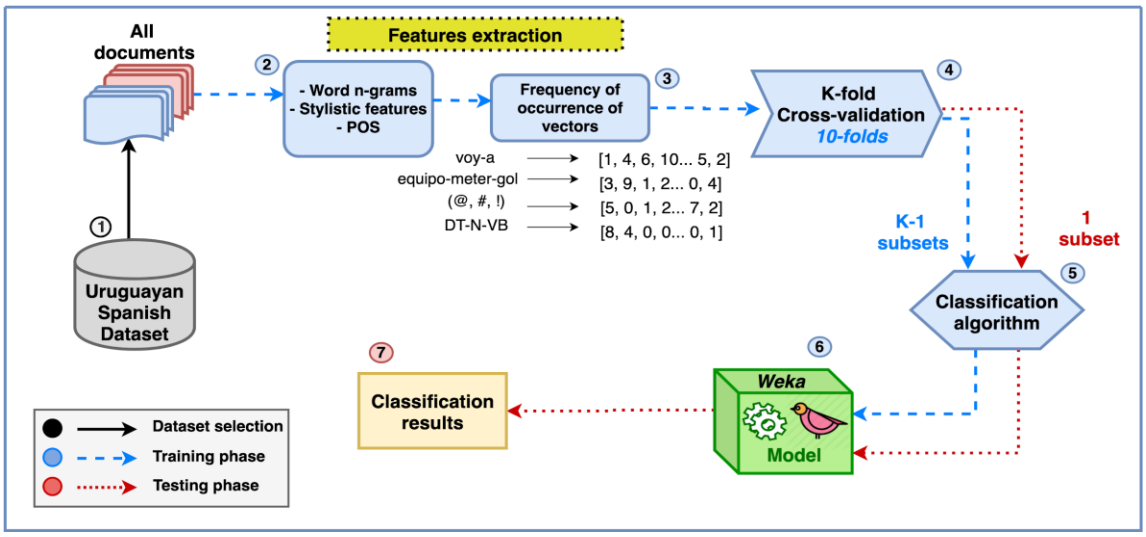

Figure 9. Sentiment classification model localization proposal.

First, all corpus' documents labeled with their polarity (1) are transformed into their representations, according to each of the four different feature types described (word n-grams, stylistic features, and POS vectors) (2, 3). Following, this data is splitted into 10 folds (4), in order to train the classification algorithms with k-1 subsets and use the remaining subset for testing (5). This way, by means of Weka, a model capable of predicting the polarity (positive, negative or neutral) associated to a text is created (6). The model is averaged against each of the folds and the result of the testing is the corresponding polarity label (7) associated with each document, which allows to evaluate the performance of the model.

[10] https://www.cs.waikato.ac.nz/ml/weka/

Evaluation and Results

Finally, this section discusses the evaluation of the localized sentiment classification model approaches, adapted to specifically handle Rioplatense Spanish. The proposed system would classify students' texts generated from their interactions in educational platforms into three sentiment classes: negative, positive or neutral polarities. Therefore, chance would be 33% of accuracy.

With the goal to assess the proposed approaches, I applied several machine learning algorithms to test the models; such as support vector machines SVM and SMO (Sequential Minimal Optimization algorithm), Naïve Bayes probabilistic algorithms, logistic regression, and decision-tree classifiers like J48. Accordingly, the testing results were reported by calculating different metrics that reflect the exactitude of the predicted classification results, including the accuracy metric (the percentage of correctly classified predictions) to allow comparison with the DIIA model, which obtained a low model accuracy of 0.472. Moreover, I obtained the F-measure metric, the harmonic mean of the complementary evaluation metrics of precision and recall (Giachanou & Crestani, 2016). Precision indicates the relationship between the number of samples correctly classified as belonging to a class and all samples that were classified as belonging to that same class. On the other hand, recall measures the relationship between the number of samples correctly classified as belonging to a class and the total number of samples of that class (Giachanou & Crestani, 2016). Following, the performance of the proposed approaches is presented, which use the Uruguayan Spanish dataset and different classification algorithms to build the models.

Original DIIA feature engineering approach. The first approach replicated the features and representations used for the DIIA sentiment classifier model. These were the most frequent 150 word trigrams in a vector space representation of their frequency of occurrence. The results of the model using the classification algorithms SMO, Naïve Bayes, the decision tree J48, and Simple Logistic are summarized in Table 4. As it can be seen, all the classifiers except for Naïve Bayes obtained a higher accuracy value than the baseline.

DIIA Features Approach		
	Accuracy	*F-Measure*
Simple Logistic	0.519	0.429
J48	0.509	0.409
SMO	0.508	0.419
DIIA Model (SVM)	**0.472**	**-**
Naïve Bayes	0.468	0.447

Table 4. Original DIIA feature engineering model evaluation results.

***Most frequent content words approach*.** The second approach involved representing the documents as vectors of the 50 most frequent content words of the collection. The method made use of the classification algorithms SMO, Naïve Bayes, the decision tree J48, and Simple Logistic, and reached higher accuracy levels than the baseline in all cases (see Table 5).

Content Words Approach		
	Accuracy	*F-Measure*
SMO	0.560	0.497
Simple Logistic	0.552	0.492
J48	0.542	0.459
Naïve Bayes	0.528	0.472
DIIA Model (SVM)	**0.472**	**-**

Table 5. Most frequent content words model evaluation results.

Stylistic features approach. This approach uses the frequencies of the stylistic elements of punctuation, user mentions, hashtags and URLs; of definite articles, and the frequency of conjunctions as features in a vector space representation. The model was built with the classification algorithms SMO, Naïve Bayes, the decision tree J48, and Simple Logistic and the results of their evaluations are presented in Table 6. All the configurations surpassed the baseline accuracy result.

Stylistic Features Approach		
	Accuracy	*F-Measure*
Simple Logistic	0.568	0.523
Naïve Bayes	0.559	0.532
SMO	0.552	0.489
J48	0.537	0.516
DIIA Model (SVM)	**0.472**	-

Table 6. Stylistic features model evaluation results.

Part-Of-Speech (POS) approach. The last approach involved the representation of the documents by their POS, having the first-person pronoun, pronominal verb, adjective, and adverb grammatical categories as features. The results of the model using the classification algorithms SMO, Naïve Bayes, the decision tree J48, and Simple Logistic are presented below (Table 7). Although the accuracy levels obtained by the model are low, they are higher than the baseline in every instance.

Part-Of-Speech (POS) Approach		
	Accuracy	*F-Measure*
Simple Logistic	0.499	0.407
J48	0.497	0.398
Naïve Bayes	0.494	0.410
SMO	0.485	0.331
DIIA Model (SVM)	**0.472**	-

Table 7. Part-Of-Speech (POS) model evaluation results.

I designed the four proposed approaches according to the rationale discussed in the previous section. As previously explained, I had the goal to localize the model to the Rioplatense Spanish dialect and, therefore, to adapt it to its implementation context in the Uruguayan educational system.

The results of the model evaluations show that all the approaches outperformed the original sentiment classifier, according to the accuracy values reached. This was true in all the iterations of each model, except for the first approach, which exceeded the baseline level in three out of four experiments. To better illustrate these outcomes, the highest accuracy values achieved for each approach are summarized in Table 8 below. As it can be seen, the Stylistic Features approach obtained the best results of the evaluations.

Model Evaluation Results		
Model	*Best Classification Algorithm*	*Accuracy*
Stylistic Features Approach	Simple Logistic	0.568
Content Words Approach	SMO	0.560
DIIA Features Approach	Simple Logistic	0.519
POS Approach	Simple Logistic	0.499
DIIA Model	**SVM**	**0.472**

Table 8. Model evaluation results summary.

7. Discussion

As it was shown by the evaluation results, the original DIIA model was outperformed by my four proposed localization approaches. Therefore, I argue that the training of the model on a Uruguayan Spanish dataset allows the representation of the linguistic characteristics of the dialect, unlike the original generic or international Spanish corpus. Moreover, the diversity of features and textual representations of the documents definitely contributed to integrally

capture these linguistic attributes and the characteristic elements of the educational and social platforms texts. The four approaches proposed included vector space representations of the most frequent word trigrams overall in the collection, the most frequent content words, frequencies of diverse stylistic elements, and different grammatical categories or POS. Furthermore, these models were built using a variety of classification algorithms: SMO, Naïve Bayes, the decision tree J48, and Simple Logistic. Consequently, the use of different classification algorithms must also have contributed to the increased accuracy values obtained.

Lastly, the evaluation of my model revealed that the stylistic features' approach reached the highest accuracy level among all proposed approaches and outperformed the original DIIA model by almost ten percentage points. The linguistic interpretation of these results is that the stylistic and format elements of the documents accurately characterize writing styles, as it has been shown in the literature (Laboreiro, Sarmento and Oliveira, 2011), but also convey emotional information. I argue that in the context of training and implementation of this model, namely in microblogging and social networking sites, users make special use of stylistic elements such as conjunctions and articles, as well as punctuation, URLs, user mentions and hashtags to express different feelings and opinions.

8. Conclusions

In line with the goals of the socio-educational Uruguayan project Plan Ceibal to foster digital inclusion and equal opportunities by means of technology, the DIIA project (Discovery of Interactions that Impact in Learning) set forth a software service for the discovery of semantic patterns that have an impact in learning, based on students' interaction in social learning

networks. The DIIA initiative included the development of a sentiment classifier that would predict the positive, negative or neutral polarity of students' texts generated in the learning platforms, such as comments, posts and messages. Accordingly, this service aimed at offering meaningful insights into students' educational experiences.

The present study focused on the sentiment analysis component of the DIIA platform, the construction of the sentiment classification model was discussed and a proposal for its improvement to reach the baseline levels achieved by state-of-the-art techniques for the sentiment classification task in Spanish was made. The hypothesis that supports the proposal is that by localizing the generic sentiment analysis module to specifically handle Rioplatense Spanish, the implementation of the classifier would offer more accurate and meaningful information about the learning experience of students in the Uruguayan educational context.

The backbone of the localization proposal was to train the classifier on a Rioplatense Spanish dataset to allow the representation of the linguistic characteristics of the Uruguayan dialect. Moreover, to integrally capture these linguistic features and the characteristic elements of the educational and social platforms texts', I proposed four model localization approaches. These explored several features and representations, including vector space representations of the most frequent word trigrams overall, the most frequent content words, frequencies of diverse stylistic elements, and different grammatical categories or Parts-Of-Speech (POS). I built these models using the classification algorithms SMO, Naïve Bayes, the decision tree J48, and Simple Logistic. After testing, it was determined that all the approaches outperformed the original sentiment classifier according to the accuracy values reached, with the stylistic features' approach obtaining the best results with the logistic

regression algorithm. Drawing from these outcomes, it can be concluded that linguistic variation is a phenomenon that definitely affects sentiment classification and thus it should be considered to efficiently tackle this and other sentiment analysis and NLP tasks. Hence, I urge the importance of compiling and working on language- and dialect-specific datasets upon the NLP research community, because the performance of supervised learning models depends on the corpus used in their training.

Furthermore, it is important to mention that advanced experimentation environments such as Weka allow language experts without a necessarily strong computational background to explore NLP and ML techniques, without requiring implementation. In the case of this research, the cross-validation procedures and the classification itself were done by means of this data mining software, and I highly recommend the use of these open access tools to my fellow linguists to venture into the area of computational linguistics.

Finally, this study establishes the grounds for further research on the potential of localizing a generic sentiment classifier in order to improve it, and, furthermore, highlights the crucial role of language in social learning analytics. Language is one of the main tools for knowledge construction and negotiation, and is a window into the complex and dynamic learning experiences of students, when analyzed properly.

9. Future Work

The evaluation of the proposed models and the outcomes achieved shed light on the possibilities of localizing a generic sentiment classifier and its potential effects for the improvement of the model. This line of research continues in favor of refining the localized

sentiment classification model, keeping in mind the growing complexity of the DIIA project. Ongoing and future work includes the following actions in order to further improve the results presented:

- Experimenting with different dataset partition, for example, using 80% of the dataset for training and the remaining 20% for testing.
- Exploring different supervised and unsupervised machine learning algorithms, such as Deep Learning (Goodfellow et al., 2016).
- Examining different features and representations for the creation of the model. For example, diverse stylistic features such as the presence of emojis and word casing; or a more detailed parsing to obtain morphosyntactic information such as appreciative prefixes.
- Compiling a new, larger and more specific dataset; a corpus composed of texts generated by Uruguayan students through their interactions with and in formal and informal educational platforms, namely CREA 2 and Facebook, in order to represent the linguistic characteristics of the Uruguayan dialect and of written online communication in educational platforms and social networks.
- Testing the proposed and forthcoming approaches in the ad hoc dataset.
- Extending the scope of the research to use fine-grained sentiment analysis systems to detect possible risk situations such as bullying, low self-esteem, isolation, and sexual harassment that may be found in students' online interaction.

10. Acknowledgements

On behalf of the DIIA team I would like to thank the support provided by the Sectoral Fund for "Digital Inclusion: Education with New Horizons" (2016) of Uruguay's National Agency for Research and Innovation (ANII) through the project FSED2_2016_1_130712.

I would like to express my appreciation to my colleagues and professors at the Universidad de las Américas Puebla and the Universidad de la República for their valuable contributions in their respective fields of expertise. I feel honored to have worked by your side and to have taken part of our international and interdisciplinary team. I thank Dr. Ofelia Cervantes Villagómez and Dr. Antonio Rico Sulayes for their guidance and support in carrying out this research. Finally, I would like to make a special mention to Dr. Esteban Castillo Juarez for his valuable teachings, mentoring, and friendship.

11. References

Aisopos, F., Papadakis, G., Varvarigou, T. (2011). Sentiment Analysis of social media content using n-gram graphs. In *Proceedings of the 3rd International Workshop on Social Media, 1*, Arizona, USA, pp. 9-14. doi:10.1145/2072609.2072614.

Al-Augby, S., Al-musawi, N. & Mezher, A. A. H. (2018). Stock Market Prediction Using Sentiment Analysis Based on Social Network: Analytical Study. *Journal of Engineering and Applied Sciences, 13*, pp. 2388-2402. DOI: 10.3923/jeasci.2018.2388.2402

Alpaydin, E. (2016). *Machine learning: the new AI*. Cambridge, Massachusetts: The MIT Press.

Behares, L. (2007). Portugués del Uruguay y educación fronteriza. In C. Brovetto, J. Geymonat y N. Brian (eds.), *Portugués del Uruguay y educación bilingüe*, pp. 99-171. Montevideo, Uruguay: Administración Nacional de Educación Pública.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*. OReilly Media, USA.

Blackmore, C. (Ed.). (2010). *Social learning systems and communities of practice*. London, the United Kingdom: Springer.

Buckingham Shum, S., & Ferguson, R. (2012). *Social Learning Analytics. Educational Technology & Society*, *15*(3), 3–26.

Carvalho, A. M. (2003). Rumo a uma definiçao do portugués uruguaio. *Revista Internacional de Lingüística Iberoamericana,1(2)*, 125-149.

Chavarría, D. (2002). *El rojo en la pluma del loro*. Barcelona: Random House Mondadori. Retrieved from: Real Academia Española: Banco de datos (CREA) [online]. Corpus de referencia del español actual. <http://www.rae.es> [April 4th, 2019].

Chen, H., Liu, X., Yin, D., Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter, 19*(2), pp. 25–35. https://doi.org/10.1145/3166054.3166058.

Chen, R. C., Huang, Y. H., Bau, C. T. & Chen, S. M. (2012). A recommendation system based on domain ontology and swrl for anti-diabetic drugs selection. *Expert Systems with Applications 39*(4), pp. 3995–4006.

Chen, X., Vorvoreanu, M. & Madhavan, K. (2014). Mining Social Media Data for Understanding Students' Learning Experiences. *IEEE Transactions on Learning Technologies, 7*(3), 246–259.

Clow, D., Ferguson, R. & Brasher, A. (2015). *LACE SoLAR Flare (OU, 2015)*. Retrieved March 10th, 2019 from https://solaresearch.org/events/flare/lace-solar-flare-ou-2015/

Corver, N., & van Riemsdijk, H. (Eds.). (2013). *Semi-lexical categories: the function of content words and the content of function words (Vol. 59)*. Walter de Gruyter.

Dabazies, A. (2003). Guambia. *Suplemento de Humor del diario Últimas Noticias*. Montevideo, Uruguay. Retrieved from: Real Academia Española: Banco de datos (CREA) [online]. Corpus de referencia del español actual. <http://www.rae.es> [April 4th, 2019].

De Melo, G., Machado, A., Miranda A. & Viera. M. (2013). *Profundizando en los efectos del Plan Ceibal. Mexico*. Instituto de Economía (UdelaR) and Centro de Investigación y Docencia Económicas (CIDE). Retreived March 10th, 2019 from: http://matchtec.mx/descargas/Plan%20Ceibal.pdf

Deng, S., Sinha, A. P., Zhao, H. (2016). Adapting sentiment lexicons to domain-specific social media texts. *Decision Support Systems*. doi:10.1016/j.dss.2016.11.001.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M. & Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers). Association for Computational Linguistics, Stroudsburg, PA, pp. 49–54.

Dubiau, L., Ale, J.M. (2013) Análisis de sentimientos sobre un corpus en espanol: experimentación con un caso de estudio. In *Proceedings of the 14th Argentine Symposium on Artificial Intelligence (ASAI)*, pp. 36–47.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (eds.). (2019). *Ethnologue: Languages of the World (22nd Edition)*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

El-Din, D. (2017). Negative Polarity Levels for Sentiment Analysis. *Data Mining And Knowledge Engineering, 9*(1), 24-28.

Elizaincín, A. (2009). Geolingüística, sustrato y contacto lingüístico: español, portugués e italiano en uruguay. *ROSAE – Congresso em Homenagem a Rosa Virgínia Mattos e Silva, 2009*.

Erden, F., Velipasalar, S., Alkar A. Z. & Cetin, A. E. (2016). Sensors in Assisted Living: A survey of signal and image processing methods. In *IEEE Signal Processing Magazine, 33*(2), pp. 36-44. doi: 10.1109/MSP.2015.2489978

Ferguson, R. (2014). Learning analytics FAQs. [Slideshare]. *The Summer Conference of The Society of College, National and University Libraries (SCONUL 2014)*. UK: Glasgow. http://www.slideshare.net/R3beccaF/learning-analytics-fa-qs

Ferrero, T., Rodríguez, C., Techera, B., & Motz, R. (2017). Analítica del aprendizaje orientada a los profesores. In *Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (WCBIE 2017)*, pp. 744–753. DOI: 10.5753/cbie.wcbie.2017.744

Fullan, M., Watson, N., & Anderson, S. (2013). *Ceibal: Next steps*. Toronto, CA: Michael Fullan Enterprises.

Galeano, E. (1979). *Días y noches de amor y de guerra*. Barcelona: Laia. Retrieved from: Real Academia Española: Banco de datos (CREA) [online]. Corpus de referencia del español actual. <http://www.rae.es> [April 4th, 2019].

Gaspari, F. Almaghout, H. & Doherty, S. (2015) A survey of machine translation competences: Insights for translation technology educators and practitioners. *Perspectives, 23*(3), pp. 333-358, DOI: 10.1080/0907676X.2014.979842

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. USA: O'Reilly Media.

Giachanou, A. & Crestani, F. (2016). Like it or not: A survey of Twitter sentiment analysis methods. *ACM Computing Surveys (CSUR) 49*(2), Article 28. DOI: http://dx.doi.org/10.1145/2938640

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press.

Gothelf, J. & Seiden, J. (2016). *Lean UX: Applying Lean Principles to Improve User Experience (2nd Ed.)*. Sebastopol, CA: O'Reilly.

Hajmohammadi, M. S., Ibrahim, R., & Ali Othman, Z. (2012). Opinion Mining and Sentiment Analysis: A Survey. *International Journal of Computers & Technology*, *2*(3), pp. 171-178.

Hamdan, H., Bechet, F. & Bellot, P. (2013). Experiments with DBpedia, WordNet and ´ SentiWordNet as resources for sentiment analysis in micro-blogging. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13), Vol. 2*. Association for Computational Linguistics, pp. 455–459.

Hamid, O. H., Smith, N. L. & Barzanji, A. (2017). Automation, per se, is not job elimination: How artificial intelligence forwards cooperative human-machine coexistence. *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)*, Emden, pp. 899-904. doi: 10.1109/INDIN.2017.8104891

Harris, S. C., Zheng, L., Kumar, V. & Kinshuk. (2014). Multi-dimensional Sentiment Classification in Online Learning Environment. *2014 IEEE Sixth International Conference on Technology for Education*, Clappana, pp. 172-175. doi: 10.1109/T4E.2014.50

Hladka, B., Holub, M. (2015). A Gentle Introduction to Machine Learning for Natural Language Processing: How to Start in 16 Practical Steps. *Language and Linguistics Compass, 9*, pp. 55 - 76. doi:10.1111/lnc3.12123

Holmes, J. (2012). An Introduction to Sociolinguistics. Harlow, UK: Pearson.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. (2018). Artificial intelligence in radiology. *Nature reviews. Cancer, 18*(8), pp. 500–510. doi:10.1038/s41568-018-0016-5

Hussein, D. M. E.-D. M. (2016). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*. doi:10.1016/j.jksues.2016.04.002

Instituto Cervantes. (2018). *El español: una lengua viva. Informe 2018*. https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_viva_2018.pdf

Jindal, N. and Liu, B. (2008) Opinion spam and analysis. In *Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008)*. doi:10.1145/1341531.1341560

Johnson, L., Adams Becker, S., Cummins, M., Estrada, V., Freeman, A., & Hall, C. (2016). *NMC Horizon Report: 2016 Higher Education Edition*. Austin, Texas, USA: The New Media Consortium.

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing: An Introduction to Language Natural Processing, Computational Linguistics, and Speech Recognition (2nd ed.)*. Upper-Saddle River, NJ: Pearson-Prentice Hall.

Kagklis, V., Karatrantou, A., Tantoula, M., Panagiotakopoulos, C. T., & Verykios, V. S. (2015). A Learning Analytics Methodology for Detecting Sentiment in Student Fora:

A Case Study in Distance Education. *European Journal of Open, Distance and E-Learning, 18*(2), 74-94. doi: https://doi.org/10.1515/eurodl-2015-0014

Kamal, A. (2015). Review Mining for Feature-Based Opinion Summarization and Visualization. *International Journal of Computer Applications, 119*(17). 10.5120/21157-4183.

Kaplan, A. & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons, 62*(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004.

Kaur, H., Mangat, V. and Nidhi, "A survey of sentiment analysis techniques," 2017 *International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, India, pp. 921-925. doi: 10.1109/I-SMAC.2017.8058315

Khan, F. H., Bashir, S. & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems, 57*, pp. 245–257. DOI:http://dx.doi.org/10.1016/j.dss.2013.09.004

Kharde, V. A. & Sonawane, S.S. (2016). Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications, 139*(11), pp. 5–15.

Khuc, V. N., Shivade, C., Ramnath, R. & Ramanathan, J. (2012). Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC'12)*. ACM, New York, NY. DOI:http://dx.doi.org/ 10.1145/2245276.2245364

Kiritchenko, S., Zhu, X. & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research, 50*, pp. 723–762.

Kumar, R., Ojha, A. K., Malmasi, S. & Zampieri, M. (2018). Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC)*. pp. 1-11.

Laboreiro, G., Sarmento, L., & Oliveira, E. (2011). Identifying automatic posting systems in microblogs. *Progress in Artificial Intelligence Conference*. Lisbon, Portugal.

Le Thi, T., Quan, T. T. & Phan Thi, T. (2017). Ontology-Based Entity Coreference Resolution for Sentiment Analysis. In *Proceedings of the Eighth International Symposium on Information and Communication Technology (SoICT 2017)*, pp. 50-56. ACM, New York, NY, USA. DOI: https://doi.org/10.1145/3155133.3155168

Lindquist, H. (2009). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University.

Liu, B. & Zhang, L. (2012). A Survey of Opinion Mining and Sentiment Analysis. In Aggarwal, C. C., Zhai, C.X., (eds.) *Mining Text Data*, pp. 415-463. Kluwer Academic Publishers.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Liu, B. (2015). *Sentiment analysis: mining opinions, sentiments, and emotions*. New York, NY, USA: Cambridge University Press.

Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion, 36*, pp. 149–161. doi:10.1016/j.inffus.2016.11.012

Long, P., & Siemens, G. (2011). Penetrating the fog: Analytics in learning and education. *EDUCAUSE Review, 46*(5), 31-40. Retrieved March 8th, 2019 from http://www.educause.edu/ero/article/penetrating-fog-analytics-learning-and-education

Manzoor Hakak, N., Mohd, M., Kirmani, M. & Mohd, M. (2017). Emotion analysis: A survey. *2017 International Conference on Computer, Communications and Electronics (Comptelix)*, Jaipur, pp. 397-402. doi: 10.1109/COMPTELIX.2017.8004002.

Martínez Cámara, E. (2015). *Sentiment Analysis in Spanish* [Doctoral Thesis]. ISBN 978-84-16819-02-7. Retrieved March 10th, 2019 from: ruja.ujaen.es/bitstream/10953/727/1/9788416819027.pdf

Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A., & Mitkov, R. (2015). Polarity classification for Spanish tweets using the COST corpus. *Journal of Information Science, 41*(3), 263–272. https://doi.org/10.1177/0165551514566564

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), pp. 1093–1113. doi:10.1016/j.asej.2014.04.011

Mejova, Y. & Srinivasan, P. (2011). Exploring Feature Definition and Selection for Sentiment Classifiers. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011).*

Meo, R. & Sulis, E. (2017). Processing Affect in Social Media: a comparison of methods to distinguish emotions in Tweets. *ACM Transactions on Internet Technology*, *17*(1).

Merriam-Webster. (2019). Diminutive. Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/diminutive

Mohammad, S. M., Kiritchenko, S. & Zhu, X. (2013). NRC-Canada: Building the state-of the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, pp. 321–327.

Montoyo, A., Martínez-Barco, P., & Balahur, A. (2012). Subjectivity and sentiment analysis. *An overview of the current state of the area and envisaged developments. Decision Support Systems*, *53*(4), 675–679. doi:10.1016/j.dss.2012.05.022

Mori, M., Tambucho M. & Cardozo, D. (2016). *Estudio de reputación a partir de comentarios extraídos de redes sociales* [Thesis]. Retrieved on March 2nd, 2019 from: http://www.fing.edu.uy/inco/grupos/pln/prygrado/InformeReputacion.pdf

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: The MIT Press.

Nenkova A., McKeown K. (2012) A Survey of Text Summarization Techniques. In: Aggarwal C., Zhai C. (eds) *Mining Text Data. Springer*, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4_3

Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2010). Affect Analysis Model: novel rule-based approach to affect sensing from text. *Natural Language Engineering*, *17*(1), pp. 95–135. doi:10.1017/s1351324910000239

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS one*, *10*(12), pp. 1-22. doi: 10.1371/journal.pone.0144296

Ortega-Mendoza, R.M. & López-Monroy, A.P. (2018) The winning approach for author profiling of mexican users in twitter at mex.a3t@ibereval-2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*. CEUR WS Proceedings.

Ortega, R., Fonseca, A. & Montoyo, A. (2013). SSA-UO: Unsupervised twitter sentiment analysis. In *Proceedings of the 7th International Workshop on Semantic Evaluation - 2nd Joint Conference on Lexical and Computational Semantics (SemEval'13)*. Association for Computational Linguistics, pp. 501– 507.

Ouertatania, A., Gasmib, G. & Latiri, C. (2018) Argued opinion extraction from festivals and cultural events on Twitter. *International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2018)*, Belgrade, Serbia.

Palacios, A. (2015). Dialectos del español de América: Chile, Rio de la Plata y Paraguay. In J. Gutiérrez-Rexach (ed.), *Enciclopedia de Lingüística Hispánica*, pp. 330-340. London: Routledge.

Pang B. & Lee L. (2009). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2*(1-2), pp. 1–135.

Peng, Q. & Zhong, M. (2014). Detecting Spam Review through Sentiment Analysis. *Journal of Software, 9*(8), pp. 2065-2072. doi: 10.4304/jsw.9.8.2065-2072.

Plan Ceibal (2017). Plan Ceibal, 10 años. *Gerencia de Comunicación de Plan Ceibal*. Retrieved March 10th, 2019 from: https://www.ceibal.edu.uy/storage/app/media/documentos/ceibal-10-2.pdf

Plaza Noblía, H. (1991). *La cerrazón*. Montevideo, Uruguay: Instituto Nacional del Libro. Retrieved from: Real Academia Española: Banco de datos (CREA) [online]. Corpus de referencia del español actual. <http://www.rae.es> [April 4th, 2019].

Prabowo, R. & Thelwall, M. (2009). Sentiment analysis: *A combined approach. Journal of Informetrics, 3*, pp. 143-157, 2009.

Ramteke, J. Shah, S., Godhia, D. & Shaikh, A. (2016). Election result prediction using Twitter sentiment analysis. *2016 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, pp. 1-5. doi: 10.1109/INVENTIVE.2016.7823280

Real Academia Española (2009). Voseo. *Nueva gramática de la lengua española*.

Real Academia Española (2009). Che. *Diccionario de la Real Academia Española*. Retrieved on April 4th, 2019 from: https://dle.rae.es/?id=8gJicRh|8gKEtDx.

Saif, H., He, Y., Fernandez, M. & Alani, H. (2016). Contextual semantics for sentiment analysis of twitter. *Information Processing & Management, 52*(1), pp. 5–19. DOI:http://dx.doi.org/10.1016/ j.ipm.2015.01.005

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Schmid, H. (n. d.). Spanish tagset documentation. *TreeTagger - a part-of-speech tagger for many languages.* Retrieved May 29th , 2019 from https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt

Seifollahi, S. & Shajari, M. (2019). *Journal of Intelligent Information Systems, 52*(1), pp. 57-83. https://doi.org/10.1007/s10844-018-0504-9

Siemens, G. (2005). Connectivism: A learning theory for the digital age. *International journal of instructional technology and distance learning, 2*(1), 3–10.

Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN). (2018). *Sentiment Analysis at Tweet level*. Retrieved March 2nd, 2019 from: http://www.sepln.org/workshops/tass/2018/task-1/

Stroh, E., & Mathur, P. (2016). *Question Answering Using Deep Learning*. Retrieved on April 2nd, 2019, from: https://cs224d.stanford.edu/reports/StrohMathur.pdf

Suero Montero, C. & Suhonen, J. 2014. Emotion analysis meets learning analytics: online learner profiling beyond numerical data. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research (Koli Calling '14)*, 165-169. New York, NY, USA: ACM. DOI: https://doi.org/10.1145/2674683.2674699

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics, 37*, pp. 267-307, 2011.

Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology, 63*(1), pp. 163–173.

Tubishat, M., Idris, N., Abushariah, M. A. M. (2018). *Implicit aspect extraction in sentiment analysis: Review, taxonomy, opportunities, and open challenges. Information Processing & Management, 54*, pp. 545 - 563. doi.org/10.1016/j.ipm.2018.03.008.

Wardhaugh, R. (2010). *An introduction to sociolinguistics (6th ed.)*. New York, NY, USA: Blackwell.

Wei, L. Conceptual and Methodological Issues in Bilingualism and Multilingualism Research. (2012) In Bhatia T. K., Ritchie W. C. (eds.). *The handbook of bilingualism and multilingualism (2nd Edition)*. Wiley-Blackwell. ISBN 9781444334906.

Wells, G., & Claxton, G. (2002). Sociocultural perspectives on the future of education. In G. Wells & G. Claxton (Eds.), *Learning for Life in the 21st Century* (pp. 1-19). Oxford: Blackwell.

Wen, M., Yang, D. & Rosé, C. P. (2014a). Linguistic reflections of students' engagement in massive open online courses. *The International AAAI Conference on Web and Social Media (ICWSM 2014)*. Retrieved from http://www.cs.cmu.edu/~mwen/papers/icwsm2014-camera-ready.pdf

Wen, M., Yang, D. & Rosé, C. P. (2014b). Sentiment Analysis in MOOC Discussion Forums: What does it tell us?. *Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014)*. Retrieved from: http://www.cs.cmu.edu/~mwen/papers/edm2014-camera-ready.pdf

Yang, J.-Y., Kim, H.-J., & Lee, S.-G. (2010). Feature-based product review summarization utilizing user score. *Journal of information science and engineering, 26*, pp. 1973-1990.

Ye, Q., Zhang, Z. & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*, pp. 6527-6535, 2009

Yuan, H., Xu, W., Li, Q. & Lau, R. (2018). *Annals of Operation Research, 270*(1–2), pp 553–576. https://doi.org/10.1007/s10479-017-2421-7

Zhang, Z. Ye, Q., Zhang, Z., & Li, Y. (2011). Sentiment classification of Internet restaurant reviews written in Cantonese. *Expert Systems with Applications, 38*, pp. 7674-7682, 2011.

12. Appendix

TreeTagger's[11] Spanish Tagset (Schmid, n. d.).

ACRNM	acronym (ISO, CEI)
ADJ	Adjectives (mayores, mayor)
ADV	Adverbs (muy, demasiado, cómo)
ALFP	Plural letter of the alphabet (As/Aes, bes)
ALFS	Singular letter of the alphabet (A, b)
ART	Articles (un, las, la, unas)
BACKSLASH	backslash (\)
CARD	Cardinals
CC	Coordinating conjunction (y, o)
CCAD	Adversative coordinating conjunction (pero)
CCNEG	Negative coordinating conjunction (ni)
CM	comma (,)
CODE	Alphanumeric code
COLON	colon (:)
CQUE	que (as conjunction)
CSUBF	Subordinating conjunction that introduces finite clauses (apenas)
CSUBI	Subordinating conjunction that introduces infinite clauses (al)
CSUBX	Subordinating conjunction underspecified for subord-type (aunque)
DASH	dash (-)
DM	Demonstrative pronouns (ésas, ése, esta)
DOTS	POS tag for "..."
FO	Formula
FS	Full stop punctuation marks
INT	Interrogative pronouns (quiénes, cuántas, cuánto)
ITJN	Interjection (oh, ja)
LP	left parenthesis ("(", "[")
NC	Common nouns (mesas, mesa, libro, ordenador)
NEG	Negation
NMEA	measure noun (metros, litros)
NMON	month name
NP	Proper nouns
ORD	Ordinals (primer, primeras, primera)
PAL	Portmanteau word formed by a and el
PDEL	Portmanteau word formed by de and el
PE	Foreign word
PERCT	percent sign (%)

[11] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

PNC	Unclassified word
PPC	Clitic personal pronoun (le, les)
PPO	Possessive pronouns (mi, su, sus)
PPX	Clitics and personal pronouns (nos, me, nosotras, te, sí)
PREP	Negative preposition (sin)
PREP	Preposition
PREP/DEL	Complex preposition "después del"
QT	quotation symbol (" ' `)
QU	Quantifiers (sendas, cada)
REL	Relative pronouns (cuyas, cuyo)
RP	right parenthesis (")", "]")
SE	Se (as particle)
SEMICOLON	semicolon (;)
SLASH	slash (/)
SYM	Symbols
UMMX	measure unit (MHz, km, mA)
VCLIger	clitic gerund verb
VCLIinf	clitic infinitive verb
VCLIfin	clitic finite verb
VEadj	Verb estar. Past participle
VEfin	Verb estar. Finite
VEger	Verb estar. Gerund
VEinf	Verb estar. Infinitive
VHadj	Verb haber. Past participle
VHfin	Verb haber. Finite
VHger	Verb haber. Gerund
VHinf	Verb haber. Infinitive
VLadj	Lexical verb. Past participle
VLfin	Lexical verb. Finite
VLger	Lexical verb. Gerund
VLinf	Lexical verb. Infinitive
VMadj	Modal verb. Past participle
VMfin	Modal verb. Finite
VMger	Modal verb. Gerund
VMinf	Modal verb. Infinitive
VSadj	Verb ser. Past participle
VSfin	Verb ser. Finite
VSger	Verb ser. Gerund
VSinf	Verb ser. Infinitive

Hoja de firmas

Tesis que, para completar los requisitos del Programa de Honores presenta la estudiante **María José Díaz Torres ID: 153452**

Director de Tesis

Dra. Ofelia Delfina Cervantes Villagómez

Presidente de Tesis

Dr. Antonio Rico Sulayes

Secretario de Tesis

Dr. Esteban Castillo Juarez