Department of Management, Economics and Industrial Engineering

Hybrid Deep Learning for Sentiment Analysis and Hate Speech Detection

Ph.D. Candidate: Claudia Volpetti
Doctoral Program Coordinator: Prof. Paolo Trucco
Thesis Supervisor: Prof. Carlo Vercellis
Thesis co-Supervisor: Prof. Carlotta Orsenigo

2019 - XXXII Ph.D. Cycle

POLITECNICO
MILANO 1863

To my beloved Andrea and Micol.

Author's biography

Claudia Volpetti has a degree in Computer Science Engineering with Major in Management Engineering and two Postgraduate Diplomas in Computer Science Education and Physics at Sapienza University of Rome and Roma Tre University. In 2006, at the beginning of the Digital TV Transition, she provides, using a prototype decoder, brand new guidelines for Usability of Digital TV Interfaces vs ISO9241 Web Design guidelines by her project degree developed for Mediaset and Publitalia with Thesis Supervisor Prof.ssa Tiziana Catarci, Sapienza University of Rome. Claudia has been working as Project Manager for Siemens, Poste Italiane and CONSEL Consorzio Elis and as SMEs (Small and medium-sized enterprises) Account Manager at Hewlett-Packard Italy as responsible for the North-Central Italian Customers Portfolio.

As researcher, she previously worked as Research Fellow at the University of Rome - Sapienza on Modeling and Simulation of Complex Systems and Complex Decision Making with a particular focus on the application of System Thinking and System Dynamics (SD) methodologies to Policies Evaluation. From 2014 to 2016 she has been working as Project Manager on the ATTACS European Project. The project - funded under the Terrorism and other Security-related Risks (CIPS) EU program - aimed to build a tool to support decision making in the field of transports protection. As researcher for PERSEUS European Project (FP7 - CORDIS), in 2015 she has been working on the design of a model aimed to analyze the effectiveness of the European Union external borders management and its consequences on migration flows in the Mediterranean Sea.

She is actually working as Researcher Phd Candidate at the Machine Learning and Big Data Analytics research group at Politecnico di Milano. Her research focuses on Machine Learning and Natural Language Processing applications ranging from Sentiment Analysis to Hate Speech and Misogyny Detection. She is an Executive Education Instructor at MIP School of Management and at the Master in Business Analytics and Big Data.

She is a Diversity & Inclusion advocate and in 2018 she is co-founder with Prof. Carlotta Orsenigo of the Milan Chapter of Women in Machine Learning Data Science (WIMLDS) a community of more than 500 members. WIMLDS's mission is to support and promote women and non-binary members who are practicing, studying or interested in the fields of machine learning and data science.

Ph.D. Program

This work is Claudia Volpetti's Ph.D thesis as a result of the Ph.D program at the research group led by prof.Carlo Vercellis at the Department of Management, Economics and Industrial Engineering, at Politecnico di Milano **(Milan, Italy)** and directed by Prof.Paolo Trucco. During her Ph.D she has been working as Visiting Researcher at IDSIA - Dalle Molle Institute for Artificial Intelligence **(Lugano, Switzerland)** directed by Prof. Luca Maria Gambardella and during her PhD she collaborated with Computer Science Deptartment of the University of Milano-Bicocca.

Publications record

1. Orsenigo C., Vercellis C., and Volpetti C. *"Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis"*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Volume 11314 LNCS, 2018, Pages 567-575, Springer, DOI: 10.1007/978-3-030-03493-1_59

2. Nozza D., Volpetti C., Fersini E. *"Unintended Bias in Misogyny Detection"*, in IEEE-WIC-ACM International Conference on Web Intelligence, Thessaloniki, Greece, 2019

3. Volpetti C., Antonucci A., Kanjirangat V. *"Temporal Word Embeddings for Narrative Understanding"*, in International Conference on Machine Learning and Computing (ICMLC), Shenzhen, China, 2020

Ph.D. Dissemination Strategy

- **27th February 2018** *Title: Hybrid Deep Learning Techniques for Sentiment Analysis* - Practitioners Event Conference Speaker - AIDIVE2018 - Event organized by E4 Computer Engineering with NVIDIA, Politecnico di Milano in collaboration with IBM. www.aidive.it (Polytechnic of Milan, Italy).

- **22nd November 2018** *Title: Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis* - International Conference Speaker - IDEAL 2018: 19th International Conference on Intelligent Data Engineering and Automated Learning - aida.ii.uam.es/ideal2018 - Session Natural Language Processing Computational Linguistics. (Autonomous University of Madrid, Spain).

- **6 March 2019** *Title: Fairness and Bias in Artificial Intelligence* - Practitioners Event Conference Speaker - WIDS Conference at Ernst & Young. Event organized by Ernst & Young in collaboration with RLadies to inspire educate data scientists regardless of gender and support women in the field. widsmilan.com (EY Milan, Italy).

- **16th May 2019** *Title: A Study on Biased Words for Hate Speech Against Women* - Practitioners Event Conference Speaker - Women in Machine Learning and Data Science (WIMLDS) Workshop at Google Italy. Co-speaker Debora Nozza, Post-Doc Research Fellow at University of Milano-Bicocca.

- **16 October 2019** *Title: Unintended Bias for Misogyny Detection* - International Conference Speaker - EEE WIC ACM International Conference on Web Intelligence. Session Web of People - Social vulnerabilities and tendencies. webintelligence2019.com (Thessaloniki, Greece)

Related Projects as Diversity and Inclusion in Data Science Advocate

Co-founder with Prof.Carlotta Orsenigo of Milan Chapter (WiMLDS) Women in Machine Learning Data Science gathering 500 data scientists in Milan area. WiMLDS mission is to support and promote women and gender minorities who are practicing, studying or are interested in the fields of machine learning and data science. We create opportunities for members to engage in technical and professional conversations in a positive, supportive environment by hosting talks by women and gender minority individuals working in data science or machine learning, as well as hosting technical workshops, networking events and hackathons. We are inclusive to anyone who supports our cause regardless of gender identity or technical background. http://wimlds.org/

Abstract

The aim of this thesis is to deploy efficient *algorithms* to automatically understand *online user-generated discussions*. In recent years, governments worldwide supported by the increasing media pressure and recent serious crime events, are demanding that social media companies, online companies, media platforms and related private stakeholders take more responsibility for what appears in their virtual spaces and are asking them to invest more in the *early detection of users emotions* (especially negative) and *fast removal of hostile and hateful contents*. Consequently, this pressure is resulting in higher companies' research investments on *efficient* algorithms for a newly born Natural Language Processing task but still with very limited research literature available, namely *Hate Speech Detection*. On the other hand, with the growing interest in ethics and sustainability issues, efficiency of Hate Speech Detection algorithms is measured lately also in terms of the *biases* affecting the algorithm. *Unbiased algorithms* are models where every group (underrepresented or protected) is fairly treated by automatic systems. Finally, an increased awareness on gaining *social behaviors insights* behind hateful users comments is demanded by public authorities in order to be proactive and anticipate violent online events. In this multifaceted context, this thesis advocate the use of *Deep Learning* methods as an efficient approach to reach *faster*, *accurate*, *unbiased* and *aware* algorithms for Hate Speech Detection by working on three different specific domains of application. Firstly, this work will introduce and implement *new hybrid representations of user-generated comments* for text-classification leveraging strengths of classical machine learning and deep learning techniques and outperforming previous attempts in literature. Secondly, this research will design a hate speech (specifically focused on misogyny) detection deep learning model that demonstrated to obtain the best classification performance in the state-of-the-art. In the same study, experimental results also will confirm the ability of the *bias mitigation treatment* implemented to reduce the unintended bias in online micro-blogging platforms, such as Twitter. Finally, we propose *dynamic representations of words* as a suitable deep learning tool to study the evolution of users roles and their sentiments across the plot of a narrative text or an online discourse; that could be used for the identification of victims/aggressors in Hate Speech Detection models. Thesis results are promising, and the empirical research outcomes demonstrated to support the working ideas behind this PhD work. From a *methodological point of view*, the Hate Speech Detection task will be addressed and studied by leveraging the wide literature available for the closely related and widely studied Sentiment Analysis task. Both tasks will be object of this thesis but with different approaches and scopes: (i) the Sentiment Analysis task and its wide literature will be investigated uniquely in order to retrieve state-of-the-art approaches and methodologies for text classification of sentences sentiment-wise; (ii) by leveraging the wide literature and the large amount of benchmark data sets available for Sentiment Analysis, new methodologies and techniques will be specifically designed exclusively for the Hate Speech Detection task.

Except for the first paper in the collection[1], where a new approach is tested on a Sentiment Analysis task due to a lack of Hate Speech Datasets back when the paper was written, any further analysis on Sentiment Analysis state-of-the-art methods is out of scope for this thesis. As future research developments, as second step in the direction of further leveraging the interplay between these tasks, we envision the use of Transfer Learning between Sentiment Analysis and Hate Speech Detection in order to improve the latter's performances. On the other hand, this work motivates and envisions further investigations of the use of temporal embeddings for the identification of victims and aggressors in hate speech dialogues, responding to the need of providing further steps in the direction of designing tools able to *anticipate* and *prevent* extreme incidents in online and offline spaces.

Key words

Hybrid Deep Learning; Machine Learning; Deep Learing; Sentiment Analysis; Hate Speech Detection; Hybrid Sentence Representations; Unbiased Algorithms; Temporal Word Embeddings; Narrative Understanding.

[1]Orsenigo C., Vercellis C., and Volpetti C."Concatenating or Averaging? HybridSentences Representations for Sentiment Analysis", Lecture Notes in Computer Sci-ence (including subseries Lecture Notes in Artificial Intelligence and Lecture Notesin Bioinformatics), Volume 11314 LNCS, 2018, Pages 567-575, Springer, DOI:10.1007/978-3-030-03493-159

LIST OF APPENDED PAPERS

The PhD thesis is a collection of papers. The following abstracts are from the papers that present relevant results for the PhD work.

PAPER I

Concatenating or Averaging?
Hybrid Sentences Representations for Sentiment Analysis

Carlotta Orsenigo, Carlo Vercellis, and Claudia Volpetti
***
Department of Management, Economics and Industrial Engineering,
Politecnico di Milano, Via Lambruschini 4b, 20156 Milan, Italy
carlotta.orsenigo,carlo.vercellis,claudia.volpetti@polimi.it

Abstract

Performances in sentiment analysis - the crucial task of automatically classifying the huge amount of users' opinions generated online - heavily rely on the representation used to transform words or sentences into numbers. In the field of machine learning for sentiment analysis the most common embedding is the bag of words (BOW) model, which works well in practice but which is essentially a lexical conversion. Another well-known method is the Word2vec approach which, instead, attempts to capture the meaning of the terms. Given the complementarity of the information encoded in the two models, the knowledge offered by Word2vec can be helpful to enrich the information comprised in the BOW scheme. Based on this assumption we designed and tested four hybrid sentence representations which combine the two former approaches. Experiments performed on publicly available datasets confirm the effectiveness of the hybrid embeddings which led to a stable increase in the performances across different sentiment analysis domains.

Keywords

Text classification; Sentiment analysis; Machine learning; Word vectors; Word2vec; Bag of words; Hybrid sentence representation.

PAPER II

Unintended Bias in Misogyny Detection

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini
***
University of Milano - Bicocca Milan, Italy
debora.nozza,elisabetta.fersini@unimib.it
Politecnico di Milano - Milan, Italy
claudia.volpetti@polimi.it

Abstract

During the last years, the phenomenon of hate against women increased exponentially especially in online environments such as microblogs. Although this alarming phenomenon has triggered many studies both from computational linguistic and machine learning points of view, less effort has been spent to analyze if those misogyny detection models are affected by an unintended bias. This can lead the models to associate unreasonably high misogynous scores to a non-misogynous text only because it contains certain terms, called identity terms. This work is the first attempt to address the problem of measuring and mitigating unintended bias in machine learning models trained for the misogyny detection task. We propose a novel synthetic test set that can be used as evaluation framework for measuring the unintended bias and different mitigation strategies specific for this task. Moreover, we provide a misogyny detection model that demonstrate to obtain the best classification performance in the state-of-the-art. Experimental results on recently introduced bias metrics confirm the ability of the bias mitigation treatment to reduce the unintended bias of the proposed misogyny detection model.

Keywords

Misogyny Detection; Bias Measuring; Bias Mitigation; Deep Learning.

PAPER III

Temporal Word Embeddings for Narrative Understanding

Claudia Volpetti, Vani K and Alessandro Antonucci
***
Politecnico di Milano - Milan, Italy
claudia.volpetti@polimi.it
IDSIA Lugano, Switzerland
vanik,alessandro@idsia.ch

Abstract

We propose temporal word embeddings as a suitable tool to study the evolution of characters and their sentiments across the plot of a narrative text. The dynamic evolution of instances within a narrative text is a challenging task, where complex behavioral evolutions and other characteristics specific to the narrative text need to be inferred and interpreted. While starting from an existing approach to the learning of these models, we propose an alternative initialization procedure which seems to be especially suited for the case of narrative text. As a validation benchmark, we use the Harry Potter series of books as a challenging case study for such character trait evolutions. A benchmark data set based on temporal word analogies related to the characters in the plot of the series is considered. The results are promising, and the empirical validation seems to support the working ideas behind this proposal.

Keywords

Natural Language Processing; Word Embeddings; Temporal Word Embeddings; Narrative Understanding; Character-Centric Narrative Understanding; Temporal Word Analogies.

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor Prof. Carlo Vercellis for the continuous support of my Ph.D study and related research, and for his motivation.

Besides my supervisor, I would like to thank the rest of my thesis committee: Prof. Carlotta Orsenigo, Prof. Michela Arnaboldi and Prof. Paolo Trucco, for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

My sincere thanks also goes to Prof. Luca Maria Gambardella, Prof. Marco Zaffalon, Dr. Alessandro Antonucci and Dr. Vani Kanjirangat of IDSIA - Dalle Molle Institute for Artificial Intelligence, who provided me an opportunity to join their team as visiting researcher, and who gave access to their research facilities and research community. Without they precious support it would not be possible to conduct this research.

Table of Contents

List of Tables

List of Figures

1. Introduction

Nowadays, the interest in social media analysis is growing together with the growing of *social media content* and several research efforts in the *artificial intelligence* domain have been recently devoted to design accurate *text classification algorithms* for *Sentiment Analysis*, the task aiming at automatically assigning a sentiment polarity to user-generated comments [1, 2]. Sentiment Analysis is mainly used for Brand Reputation [3] and Political Communication [4] but also e.g. to forecast the impact of news on Financial Markets [5]. **The same effort has not been recorded in exploiting artificial intelligence techniques for classification tasks closely related to sentiment analysis but with major and valuable social impact, such as *Hate Speech Detection***.

Hate Speech is the growing serious phenomenon defined by the European Union in the Framework Decision 2008/913/JHA of 28 November 2008[2], as *"all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin"*. A wider definition comes from The Encyclopaedia of the American Constitution [6] defining hate speech as *"any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic"*. In recent years, media coverage of this problem have increased along with the growing political attention on the phenomenon. Lately, institutions worldwide are demanding that Social Media companies take more responsibility for what appears in their networks and are asking them to invest more in the early detection and fast removal of hostile contents. In Europe, the No Hate Speech Movement[3] program has been largely founded by the Council of Europe and the EU regulators have been pushing social media firms to remove racist and violent posts from their platforms in a timely manner for years. Lately in 2017, the German government approved a plan in April to start imposing fines of as much as 50 million euros (59$ million) on Facebook, Twitter and others if they fail to remove hate speech and fake news posts within 24 hours after being flagged [7]. Other illegal content needs to be deleted within 7 days of reporting. But still, in Europe in more than 28% of cases, it takes more than one week on average for online platforms to take down illegal content [7]. **Adding to this, the time to detect and the manual effort of screening social contents, you can get an idea of the lateness of the actual interventions against illegal contents. In this context, it becomes clear that the interest in on-line Hate Speech detection and *particularly in its automation*, is set to increase.**

In the *artificial intelligence* domain, Hate Speech Detection - defined as the task of *performing automatically* the classification of user comments as hate speech - recently gained its own research line [8, 9]. Major conferences propose more and more every year special interest workshops on Hate Speech Detection or more generally on Abusive Language Online[4]. In recent years, among the artificial intelligence approaches, *Deep Learning* tech-

[2] http://eur-lex.europa.eu

[3] https://www.nohatespeechmovement.org/.

[4] 3rd Workshop on Abusive Language Online hosted by ACL Association of Computational Linguistics con-

niques have been reported to be the state-of-the-art in several domains. The big pay-off of deep learning is its capability of learning high-level features representation directly from raw data with small or even none hand-crafted feature engineering. This capability demonstrated to be the key to deep learning success across different tasks and domains especially *when dealing with unstructured data such as text*, speech and images. Therefore, thanks to its proven effectiveness, many recent studies have started to envision the application of these techniques also to Natural Language Processing (NLP) tasks. The application of deep learning techniques to NLP tasks have already proved to achieve many state-of-the-art results, performing better than standalone NLP.

In conclusion, this document presents the research entitled *"Hybrid Deep Learning for Sentiment Analysis and Hate Speech Detection"* aimed at providing new techniques, methodologies and artificial intelligence solutions to the task of automatically detecting abusive language from user-generated comments by leveraging the literature available for sentiments detection. In the following, Section 2 frames the context and highlights the major motivations under the study of Hate Speech Detection and open challenges; Section 3 defines the research objectives and methodology; Section 4 reports the the state of the art for the Sentiment Analysis task and highlights how this can be used for Hate Speech Detection; Section 5 illustrates the research results: since this work is a collection of papers, this section will briefly present and discuss each paper of the collection, focusing on the specific goals and main results. Finally, Section 6 reports the conclusions and a brief discussion about future research developments. In the appendices section, Annex A presents the collection of three papers produced for this PhD work.

2. Research Context

In the first part of this chapter, a brief discussion on hate speech terminology, definition and related concept is reported. Motivations behind the study of hate speech automatic detection are examined in depth in the second part of the chapter, where we go deeper in the motivations to study hate speech by discussing it from an academic and practitioners point of view.

2.1 Terminology, Definition and Related Concepts

The term hate speech is an umbrella term for many offensive user-generated comments, it is a legal term in several countries and is also the most used expression to depict the phenomenon in the media. On the other hand, the scientific community is also using other related terms to speak about the same phenomenon such as *abusive messages*, *hostile messages* and *flames* [10], *offensive language* [11], *profanity* [12], *vulgar language* and *profanity-related offensive content* [13], *othering language* [14]. However, the term *hate speech* still remains one of the most used, mainly because in literature is has been used by several seminal papers such as [15, 16, 16–18] and many others recently use this

ference 2019, https://sites.google.com/view/alw3/

term not only in Computer Science and Engineering related subjects but across disciplines. In the following of this thesis, the term *hate speech* is adopted as it can be considered the most popular term in both scientific and practitioners communities to speak about offensive user-generated comments.

Two hate speech definitions have been already cited in Section 1 and even if they may appear similar, the two examples differ in many details, such as e.g. the *incitation to violence* which is explicitly cited by the first definition and not even mentioned in [6]. This paragraph aims at clarifying which are the dimensions that should be present in a unambiguous hate speech definition. In the following, those dimensions are presented along with a comparison of the hate speech term with other related concepts for each dimension. Hate speech definition is indeed complex because of the numerous related concepts that sometimes can overlap and make difficult a clear classification between them. Several of those concepts can be found in literature: *cyberbullying* [19–21], *discrimination* [22], *toxicity* [23], *flaming* [24], *extremism* [25, 26], and *radicalization* [27]. In the following we provide a list of four dimensions to uniquely define hate speech and contextually we clarify the difference between hate speech and other related concept for each dimension respectively:

1. **Hate speech is a speech attacking specific targets and groups of people** identified on the basis of specific characteristics like religion, sexual orientation, gender, ethnic origin, etc.. According to this dimension, hate speech differs from *toxic language* (defined as "toxic comments which are rude, disrespectful or unreasonable messages that are likely to make a person to leave a discussion" [43]) and *profanity* (defined as "offensive or obscene word or phrase" [23]) since those can be also perpetrated without a specific target. Hate speech also differs from *cyberbullying* where "the aggressive and intentional act is carried out repeatedly and over time, against a victim who can not easily defend him or herself" [10] while hate speech is more general and not necessarily focused on a specific person and is more about stereotypes.

2. **Hate speech is a speech inciting violence or hate.** Very close concepts are *radicalization* and *extremism*. In order to clarify the difference between those three concepts, we need first to point out that online *radicalization* is similar to *extremism* but radical discourses are usually related a subset of topics such as terrorism, anti-black communities, or nationalism [27] while extremism can be on any ideology. However in both radical and extremist discourses you can find topics like religion and war [27], recruitment of new members, social media and institutions demonization and even persuasion[26] while hate speech don't usually touch those topics and it can be more grounded in stereotypes and hence more subtle. That means that the kind of violence incited by hate speech discourses can also be subtle as in the case of stereotypes that are gradually reinforced to such an extent that can be used to justify discrimination, violence and hate against groups of people.

3. **Hate speech is a speech aiming to attack or diminish specific groups of people**. This definition makes hate speech almost indistinguishable from *discrimination*, although the latter can be used as the basis of unfair treatment in every environment

and can also refer to discriminating behaviors while hate speech is more about discrimination through verbal means.

4. **Hate Speech is not Humour** and Humour is not Hate Speech, even if this latter can carry subtle forms of discrimination e.g. through jokes playing on stereotypes. In this work, we consider these kind of jokes as hate speech because in case of long exposure of users to them, the consequences could certainly harmful towards some groups of people that could decide to leave the conversation [28].

In conclusions, in this paragraph we used the four dimensions listed above to clarify the hate speech concept while underling its difference with other very close concepts. This analysis is well summarized by [8] which proposed a complete and unambiguous definition of hate speech that we are going to use in the following of this document:

"Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used". [8]

Now that a more refined definition of what is hate speech has been introduced, in the next subsection, we will go deeper in the motivations to study this phenomenon by discussing them both from an academic and practitioners point of view.

2.2 Why study Hate Speech

The potential impact of a research focused on automatic Hate Speech is clearly sustained by current events and daily news: European Institutions are asking for more investments and some Member States started to legislate and regulate the phenomenon. Media and legislation pressure, envision that in less than a decade, Social Media companies must be completely able to early detect and ban illegal hate content. A short list of my research project stakeholders are *public authorities in charge of cyber-security*, *police* and *justice* and *social media managers* of politicians and public figures. Several motivations are encouraging more and more researchers to focus on automatic hate speech detection, a list of the main ones is reported below:

1. **European Union Commission Directives.** Hate speech is illegal in many countries worldwide. In Europe and in the world governments and institutions are conducting several initiatives aiming at decreasing the hate speech phenomenon through legislation. As already reported in Section 1, European Union Commission recently pressured Facebook, YouTube, Twitter, and Microsoft to sign an EU hate speech code requiring to remove hateful comments in less then 24h.

2. **Automatic tools are scarce.** Automated techniques aim to automatically classify text as hate speech, making its detection easier and faster for the ones that have the

responsibility to protect the public [9, 65]. Several efforts have been reported aiming at providing a fully automatic detection of hate speech, but the tools provided are still inadequate. For a more detailed analysis, please refer to Section 4 *State of the Art*.

3. **Algorithms are biased.** Automatic detection of hate speech discourses usually is provided using NLP and deep learning methods for text classification which are performing as state-of-the-art systems but unfortunately result heavily biased towards some groups of people [29, 30]. This results in systems that can better recognize some hateful comments more than others, discriminating groups of users. Section 5.2 will go deep into the bias issue of algorithms for automatic classification.

4. **Lack of Benchmark Datasets.** Research into the field of hate speech detection suffers from a serious lack of benchmark datasets due to the fact that the hate phenomenon can be very wide and affect many groups and in many cases research cannot cover the great variety of "protected groups". Lately, academic research - especially in Italy, is focusing on a set of subtypes of hate speech, such misogynous comments, hate against migrants or LGBT community [31, 32]. But the general idea is that there is a serious need of more datasets covering other categories of potential trageted categories.

5. **Hate speech Removal.** Online companies and media platforms cannot afford to lose advertisers because of their virtual space not being safe for their users [33]. The risk of associating their brands to hateful and unsafe virtual spaces is too high, so as consequence not only Social Media companies but a wider group of companies are stakeholders of a research focusing on systems able to remove hate speech.

6. **Quality of service.** Not only Social Media companies but as already discussed a wider range of companies need to keep their virtual spaces safe. Quality of service for such companies, include not only their capabilities of removing hateful comments but also of being able to provide *unbiased* services. That means being sure that their artificial intelligence algorithms are not affected by biased trainings and that can provide fair predictions. After the work of [30] and [29] by Google researchers, now unintended bias can be measured and so platforms can be compared on their "fairness". In conclusion, the quality of the service provided nowadays is highly dependent also on the capability of companies to be able to provide debiased algorithms.

7. **Social Behavior Insights.** Finally, a part from detecting hate speech comments, in order to gain useful insights on the social behaviors of their users, several stakeholders ranging from private companies to public authorities, are in demand of systems that can help in classifying the user involved in a hateful discourse as fast as possible in an automatic way. That is the motivation behind the research line aiming at automatically identiy *victims* and *aggressors* in online discussions. Such social behavior insights can be also helpful to prevent incidents in online and offline spaces, thus providing steps in the direction of what is called *anticipatory governance*.

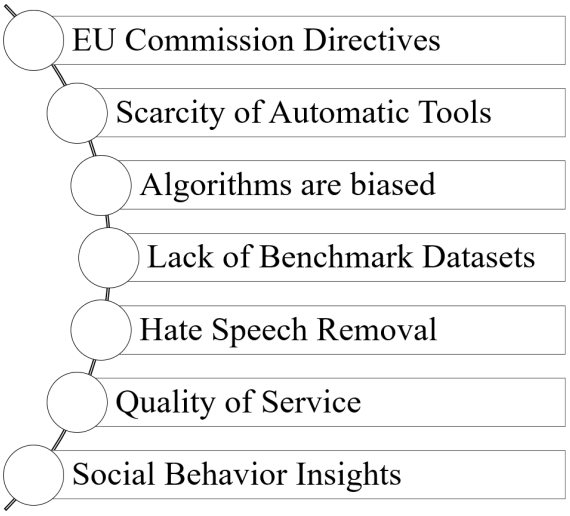

Fig. 1. Why study Automatic Hate Speech Detection.

3. Research Objectives, Structure and Methodology

This section firstly illustrate the research objectives addressed by this research along with the corresponding research questions hierarchy, thesis structure and methodology. This chapter also include a description of how the papers included in this thesis contribute to the main goal by addressing respectively each research question.

3.1 Research Objectives

Section 2.2 presented the main motivations of this work of research. Among the arguments motivating the Automatic Hate Speech Detection research line, several of them can be considered open issues still counting several stakeholders from the private and public sector of our society. Building on the research motivations and the stakeholders presented in Section 2.2, in the following, the list of open issues targeted by this thesis will be reported along with the first formulation of their corresponding research questions (see Figure 2):

1. **Compliance to Legislation**: private companies such as Facebook and Twitter have been asked to be *fast* to react in a timely manner when detecting, classifying and remove hate speech from their platforms. In order to answer readily to hate speech cases, such companies need to implement hate speech detection *automatically*, that causing a high demand of intelligent algorithms able to understand human writing and perform the text classification task autonomously without human intervention. Figure 2 shows the *Compliance to Legislation* open issue as the main cause for the high demand of Hate Speech Detection system to be *fast* and consequently to be implemented in order to be fully *automatic*.

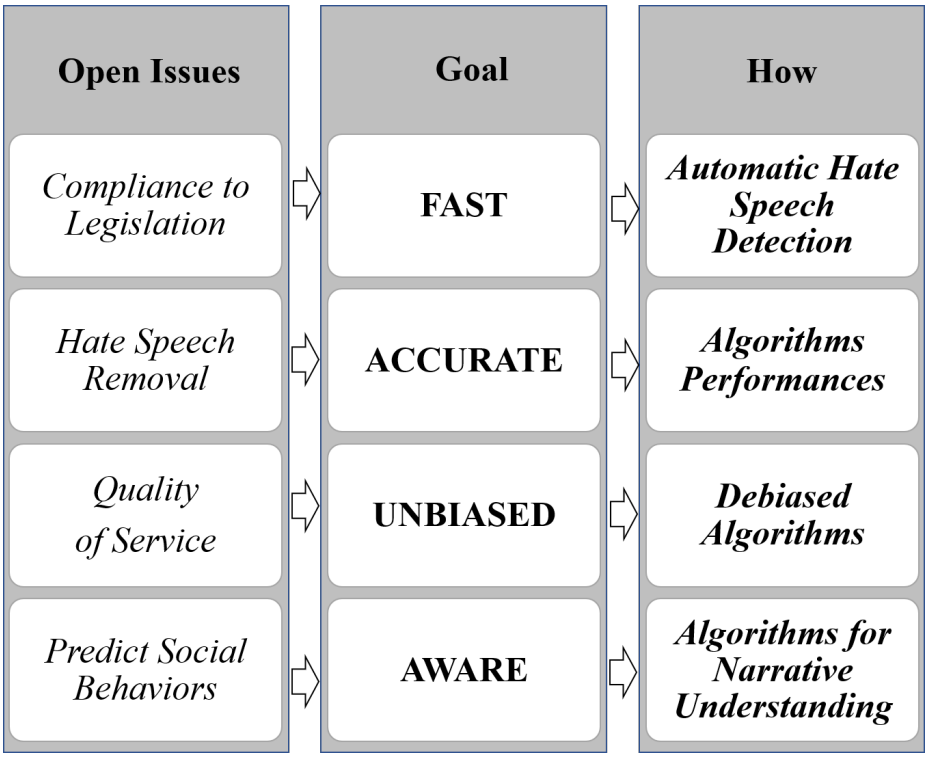

Fig. 2. Research Questions vs Open Issues Mapping.

2. **Hate Speech Removal**: not only Social Networks but many other companies need to keep their virtual spaces *safe* because they need users to spend time on their platforms and attract advertisers. Companies' virtual spaces should be hate speech free in order not to let the phenomenon to be associated to their brands, that means a high demand in *accurate* algorithms capable of classifying with high accuracy hate speech comments. On the other hand, algorithms are asked to reduce the number of false positives so that not to classify as hateful comments which are not. As shown in Figure 2, the *Hate Speech Removal* open issue leads to a high demand of *accurate* automatic systems and consequently to the need of designing *algorithms with high performances* in terms of accuracy and precision.

3. **Quality of Service**: Social media companies and other related companies providing virtual spaces for online discussions, they are in need of discouraging hate speech as far as to keep algorithms monitoring such spaces *unbiased* so that every group of people should feel comfortable at expressing opinions and communicate with other users in the platform because is treated in fair and unbiased way by the automatic systems designed. Everyone should be able to express themselves online, so we want to make conversations more inclusive. It has been revealed by [29] the serious bias of text classification algorithms when it has been demonstrated that certain systems were heavily biased against e.g. women, gay or black people. That means that e.g.

such system were biased thinking that if a comment was including the word gay that is for sure hate speech biased by a negative representation of gay people in the training data[5]. As shown in Figure 2, the *Quality of Service* open issue is recently demanding automatic systems to be more and more *unbiased* and is consequently leading to the high need of designing new *debiased algorithms* able to treat equally different groups of people participating in online discussions.

4. **Predict Social Behaviors**: the capability of gaining useful insights on social behaviors of online platforms users is stressing several stakeholders to increase the *awareness* on what is happening on the virtual spaces they provide. The challenge addressed by this research work is to automatically identify *victims* and *aggressors* in online discussions. As shown in Figure 2, the *Predict Social Behaviors* open issue leads to a high demand of *aware* platforms and consequently of *algorithms for narrative and discussion understanding*.

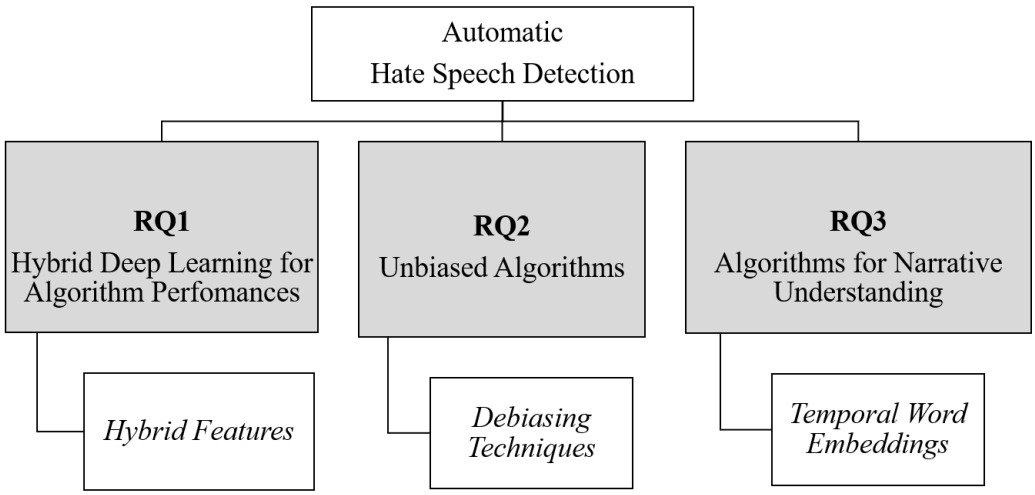

Fig. 3. Research Questions Hierarchy.

3.2 Research Questions Hierarchy and Thesis Structure

Building on the open issues reported above and the respective open questions, Figure 3 shows the research question hierarchy which is presented in detail in the following of this

[5]https://services.google.com/fb/forms/respect/

paragraph. Under the umbrella of the main goal of building *fast, accurate, unbiased and aware* hate speech detection algorithms the process of detecting hate speech, three main goals are pursued in this research work:

1. **RQ1 - Can hybrid Deep Learning techniques be effectively used in order to reach higher performances in text classification tasks?** In order to reach the goal of hate speech detection to be *accurate* (Fig. 2), *new and performing artificial intelligence* algorithms are envisioned in this work. The first research step is directed at investigating new deep learning techniques in order to reach higher performance in text classification tasks. More in detail, as it will be fully explained later in Section 5.1, by leveraging strengths of machine learning and deep learning worlds, *hybrid deep learning* architectures will be designed with a special focus on *hybrid features* representation techniques. The guess is that by working into the direction of a better understanding of the strengths and weaknesses of hybrid approaches, and by leveraging them, it will enable the design of new well-performing variants. Figure 3 shows the *RQ1 - Hybrid Deep Learning for Algorithm Performances* research question as first branch of our Research Questions Hierarchy. This research line will be the topic of the first paper named *"Concatenating or Averaging?Hybrid Sentences Representations for Sentiment Analysis"* co-authored with prof.Carlotta Orsenigo and prof.Carlo Vercellis and presented in detail in Section 5.1.

2. **RQ2 - How to measure and mitigate unintended bias in hate speech detection algorithms?** The *Quality of Service* of online platforms is an open issue that may be solved by working on debiasing algorithms. So as second step in this research work, we will investigate in designing new debiasing techniques aiming at keeping the stakeholders ' virtual spaces unbiased and fair with regards of any group of people e.g. women, migrants, LGBTQ+ community, etc. Outcomes and methodology of this second research step will be presented later in section 5.2 and are published in a paper named *"Unintened bias for Misogyny Detection"* written with Debora Nozza and Elisabetta Fersini (Università di Milano - Bicocca). In this paper, we provide a model for misogyny detection which demonstrates to obtain the best classification performance in the state-of-the-art and we address the fairness of this model by measuring and mitigating its unintended bias against women. Figure 3 shows the *RQ2 - Unbiased Algorithms* research question as the second branch of our Research Questions Hierarchy. Presented at Web Intelligence conference in October 2019 and Scopus Indexed.

3. **Q3 - Can Deep learning word representations be used in order to better understand the social interactions in virtual spaces?** Finally, as third research step, this study address the open issue related to the lack of awareness of algorithms in terms of the ability of understand and *predict social behaviors* by reading online conversations. So, the third research steps aims at leveraging deep learning representations of words in order to design algorithms able to understand the different roles that the

users play during a hate speech event. We will resort to the work done in terms if *algorithms for narrative understanding* and we advocate the use of new techniques such *temporal word embeddings* for this task. Results and methodology of this paper named "Temporal Word Embeddings for Narrative Understanding" written with Alessandro Antonucci and Vani K (IDSIA - Switzerland). Figure 3 shows the *RQ3 - Algorithms for Narrative Undestanding* research question as the second branch of our Research Questions Hierarchy.

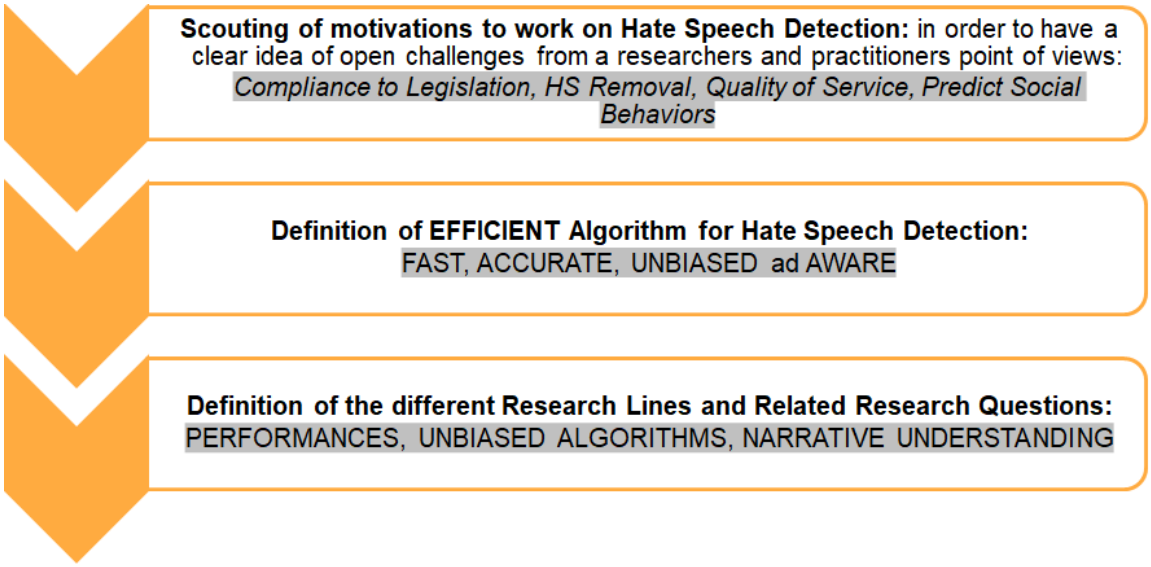

Fig. 4. Research Preliminary Steps.

3.3 Research Methodology

From a methodological point of view, in this thesis both Sentiment Analysis and Hate Speech Detection tasks are studied but with different approaches and scopes: (i) the Sentiment Analysis task and its wide literature will be investigated uniquely in order to retrieve state-of-the-art approaches and methodologies for text classification of sentences sentiment-wise; (ii) consequently, by leveraging the wide literature and the large amount of benchmark data sets available for Sentiment Analysis, new methodologies and techniques will be specifically designed exclusively for the Hate Speech Detection task. Except for the first paper in the collection[6], where a new approach is tested on a Sentiment Analysis task due to a lack of Hate Speech Datasets back when the paper was written, any further analysis on Sentiment Analysis state-of-the-art methods is out of scope for this thesis.

[6]Orsenigo C., Vercellis C., and Volpetti C."Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis", Lecture Notes in Computer Sci-ence (including subseries Lecture Notes in Artificial Intelligence and Lecture Notesin Bioinformatics), Volume 11314 LNCS, 2018, Pages 567-575, Springer, DOI:10.1007/978-3-030-03493-159

This research will be conducted on English texts since, a part from few rare exceptions, the main literature available on both tasks is on English contents. On the other hand, the main drawback of this choice is that outcomes of studies concerning English hate speech datasets don't benefit of characteristics of transferability to other languages because the culture in the home country of the language has its own weight on the classification task. Since this work is one of the first on this research topic, it has been a preferable choice though to leverage the existing literature on English texts.

The specific Research Lines and Related Research Questions have been identified by a series of research preliminary steps (Fig. 4). First, a scouting of the motivations to work on Hate Speech Detection has been carried out , in order to have a clear idea of the open challenges for the domain of interest (see Section 2). Then a definition of *efficient* Hate Speech Detection algorithm is provided in Section 3 along several dimensions: *fast, accurate, unbiased and aware*. Finally, according to these dimensions, a mapping of Research Objectives and the Research Questions Hierarchy have been designed (Section 3).

4. Sentiment Analysis - Literature Review

In this paragraph, it is reported the mapping exercise to analyze and chart the progress of research work in Deep Learning (DL) exclusively for the Sentiment Analysis task in the last few years. The focus of my work of literature review has been to map the "state-of-the-art models" from 2016 backwards. An update to more recent works has been performed recently in order to cover all academic papers published up to 2019 and it is presented in paragraph 4.3. I worked on a detailed manual analysis – aka Content Analysis - of the research publication data in order to identify: *(i)* The main models and approaches used in state-of-the-art models along with their performances in terms of accuracy, main benchmark datasets and *(ii)* The main challenges and open problems into the field of DL for Sentiment Analysis.

4.1 Search Protocol

As a preliminary phase, in order to retrieve the scientific papers related to the scope of this survey, I designed and implemented the following search protocol:

1. I firstly identified the research publication corresponding to the more recent state-of-the-art model (we will call it the root-paper) and I went through a reference search of previous state-of-the-art models.

2. I did a manual screening of the research publications referenced by the root-paper backwards, cycling the same procedure, in order to find out all those research publications that somewhere in the past demonstrated to be Deep Learning for Text Classification state-of-the-art models.

3. Inclusion Criteria: At first search, the root-paper resulted to be dated 2015, then we included by a Scopus search a further paper dated 2016 in order to cover the period of interest.

4. Exclusion Criteria: The back propagated reference search stopped to research publications dated before 2011[7].

As a result of the reference search, I obtained a number of 16 papers for a total of 50 different model variants. I did a manual cleaning of the research publications from the root-paper backwards to 2011 to find out those models that effectively used Deep Learning approaches. As results, out of the 50 models, 29 models were found to use DL algorithms (see Table 1). All the other models were identified as Machine Learning models mainly used as baseline for comparison and they were not analyzed further. This check required downloading the full text of the 16 papers from Scopus[8] and understanding the work reported to identify if the paper reports the implementation of Deep Learning models. The resulting 16 papers are listed in the references section and in Table 2. The CODE ID used in figures is straightforward: year of publication and name of the first author.

Table 1. Deep Learning Models vs other approaches.

Class	Frequency
Machine Learning	21
Deep Learnig	29

4.2 Content Analysis

The manual annotations hand-made by the author of this work, were aimed at identifying mainly models and approaches used in state-of-the-art models along with their performances in terms of accuracy, main benchmark datasets and datasets domains in the period of interest. The results of manual analysis are presented in this section. Coding fields for the Manual Reading have been fixed as follows: Class (Deep Learning or NoDL), Year, Title, Model, Family Model, Variants of previous models, Feature Representation, Insights, Issues, Main Focus.

As first tangible outcome of the manual screening, I produced a table of the **state-of-the-art models** in Deep Learning for Text Classification in Figure 2 from 2011 to 2016. For each paper, CODE ID field is a label automatically assigned to each paper composed by *year_author*, the MODEL is the name of the algorithm as provided in the paper, and the following columns are the name of the datasets on which the MODEL as been tested. A detailed description of the dataset will be provided in the following of this chapter. Performances are measured in term of percentage of accuracy over a test set. As literature

[7] 2011 has been identified by this work as the year of first broad usage of the term Deep Learning; given that we do not search behind this date.

[8] https://www.scopus.com

CODE ID	MODEL	MR	SST-1	SST-2	SUBJ	TREC	CR	MPQA
2010_NAKAGAWA[34]	Tree-CRF	77,3					81,4	86,1
2011_SILVA[35]	SVMs					95		
2011_SOCHER[36]	RAE-rand	76,8						85,7
2011_SOCHER[36]	RAE-init	77,7						86,4
2012_SOCHER[36]	MV-RNN	79						
2012_WANG[37]	MNB-uni	77,9			92,6		79,8	85,3
2012_WANG[37]	MNB-bi	79			93,6		80	86,3
2012_WANG[37]	SMV-uni	76,2			90,8		79	86,1
2012_WANG[37]	SVM-bi	77,7			91,7		80,8	86,7
2012_WANG[37]	NBSVM-uni	78,1			92,4		80,5	85,3
2012_WANG[37]	NBSVM-bi	79,4			93,2		81,8	86,3
2013_HERMANN[38]	CCAE-A	77,8						86,3
2013_HERMANN[38]	CCAE-B	77,1						87,1
2013_HERMANN[38]	CCAE-C	77,3						87,1
2013_HERMANN[38]	CCAE-D	76,7						87,2
2013_SOCHER[39]	RNTN		45,7	85,4				
2013_SOCHER[39]	RNN		43,2	82,4				
2013_SOCHER[39]	NB		41	81,8				
2013_SOCHER[39]	BiNB		41,9	83,1				
2013_SOCHER[39]	VecAvg		32,7	80,1				
2013_SOCHER[39]	SVM_SOCHER		40,7	79,4				
2013_SOCHER[39]	MV-RNN		44,4	82,9				
2013_WANG[40]	G-Dropout	79			93,4		82,1	86,1
2013_WANG[40]	F-Dropout	79,1			93,6		81,9	86,3
2014_DONG[41]	s.parser-LongMatch	78,6						85,7
2014_DONG[41]	s.parser-w/oComb	78,3						85,5
2014_DONG[41]	s.parser	79,5						86,2
2014_IRSOY_CARDIE[42]	DRNN (4,174)		49,8	86,6				
2014_KALCHBRENNER[43]	MAX-TDNN		37,4	77,1		84,4		
2014_KALCHBRENNER[43]	NBOW		42,4	80,5		88,2		
2014_KALCHBRENNER[43]	DCNN		48,5	86,8		93		
2014_KIM[44]	CNN-rand	76,1	45	82,7	89,6	91,2	79,8	83,4
2014_KIM[44]	CNN-static	81	45,5	86,8	93	92,8	84,7	89,6
2014_KIM[44]	CNN-non-static	81,5	48	87,2	93,4	93,6	84,3	89,5
2014_KIM[44]	CNN-multichannel	81,1	47,4	88,1	93,2	92,2	85	89,4
2014_LE_MIKOLOV[45]	Paragraph-Vec		48,7	87,8				
2014_YANG[46]	CRF						81,1	
2014_YANG[46]	CRF-PR-inf_lex						80,9	
2014_YANG[46]	CRF-PR-inf_disc						81,1	
2014_YANG[46]	CRF-PR-lex						81,8	
2014_YANG[46]	CRF-PR						82,7	
2015_TAI[47]	LSTM		46,4	84,9				
2015_TAI[47]	Bi-LSTM		49,1	87,5				
2015_TAI[47]	2-layer LSTM		46	86,3				
2015_TAI[47]	2-layer Bi-LSTM		48,5	87,2				
2015_TAI[47]	Dep Tree-LSTM		48,4	85,7				
2015_TAI[47]	C_Tree-LSTM - rand		43,9	82				
2015_TAI[47]	C_Tree-LSTM - GloVefix		49,7	87,5				
2015_TAI[47]	C_Tree-LSTM - GloVetun		51	88				
2016_KUMAR[48]	DMN		52.1	88.6				
2016_KUMAR[48]	DMN		52.1	88.6				

Table 2. State-of-the-art models and performances over the period from 2011 to 2016

review outcome[9], I found out that the state-of-the-art architecture in Sentiment Analysis was the LSTM (Long Short-Term Memory Networks) by [47]. Long Short-Term Memory Networks are a variant of Recurrent Neural Networks, which addresses the problem of learning long-term dependencies by introducing a memory cell that is able to preserve state over long periods of time. A far more complex multi-task architecture by [48], named DMN (Dynamic Memory Networks) also obtained state-of-the-art results on text classification for sentiment analysis in 2016. The DMN model is a potentially general architecture, belonging to a brand-new area of research of multi-task family architectures. DMN can be applied for a wide variety of NLP applications, including classification, question answering and sequence modeling.

As a further outcome of the manual content analysis, we found out that the main benchmark **datasets** used for state-of-the-art comparison are:

- **MR**: Movie reviews with one sentence per review. Classification involves detecting positive/negative reviews [49].
- **SST-1**: Stanford Sentiment Tree-bank an extension of MR but with train/Dev/test splits provided and fine grained labels (very positive, positive, neutral, negative, very negative), re-labeled by [39].
- **SST-2**: Same as SST-1 but with neutral reviews
- **Subj**: Subjectivity dataset where the task is to classify a sentence as being subjective or objective [49].
- **TREC**: TREC question dataset—task involves classifying a question into 6 question types (whether the question is about person, location, numeric information, etc.) [41].
- **CR**: Customer reviews of various products (cameras, MP3s etc.). Task is to predict positive/ negative reviews [50].
- **MPQA**: Opinion polarity detection subtask of the MPQA dataset [51]

The manual content analysis helped in framing that the state-of-the-arts models implemented in the period of analysis mainly refer to the following families of Neural Network Architectures:

- **Recursive Neural Networks**: RNN exploit the nested hierarchy and an intrinsic recursive structure of the data [36] [38];
- **Long Short Term Memory Networks**: a variant of Recurrent Neural Networks, they can learn long-term dependencies by introducing a memory cell that is able to preserve state over long periods of time. Tree-LSTM is a LSTM which exploit the dependency parsing tree of sentence [47];
- **Convolutional Neural Networks**: CNN are neural networks comprising one or more convolutional layers which are essential learning filters. CNN are mainly used to deal with images [44].

[9]Those are results from a literature review performed in 2017; a literature update is provided in the next paragraph.

- **Dynamic Memory Networks**: DMN is a neural network based framework for general question answering tasks that is trained using raw input-question-answer triplets [48].

4.3 Recent Research Advances

Ground-breaking ideas are occurring at an unprecedented pace lately in the Natural Language Processing research community and in particular from the new architectures coming from a Deep Learning approach. Christopher D. Manning, professor at Stanford and one of the major expert in NLP said this new techniques can be referred to as "Deep Learning Tsunami"[10] in Computational Linguistics. As already anticipated, deep learning has the major pay-off of learning high-level features representation directly from raw data with small or even none hand-crafted feature engineering and consequently to be outstanding in performances when dealing with unstructured data. Deep Learning also use a hierarchy of layers able to leverage the compositionality of neural networks architectures but the real game changing innovation is nowadays the use of distributed word representations such as word2vec [52] or GloVe [53]. Recurrent Neural Networks and in particular Long-Short Term Memory Networks [54] have been architectures that performed as state-of-the-art on NLP tasks and along with the distributed word vectors representations, left far behind every previous approach. In the last years, the Attention Mechanisms first introduced by [55] has gained an increased popularity and has found a broad application in a wide range of NLP tasks. In [56], authors augmented the long short-term memory (LSTM) network with a hierarchical attention mechanism consisting of a target-level attention and a sentence-level attention. One of the seminal idea of 2018, has been the Transformer architecture [57] builds on the attention mechanism and aiming at solving the long term dependencies in a more efficient way than the previous state-of-the-art architecture, the LSTM neural networks [54]. The Transformer architecture demonstated to be the state-of-the-art model in modern NLP and building on this method, Google's BERT model has been designed. BERT (Bidirectional Encoder Representations from Transformers) [58] is a recent paper by Google AI research that is performing state-of-the-art results in a broad variety of natural language processing tasks by applying a multi-layer bidirectional Transformer encoder. Lately, the use of capsule network - initially introduced for image classification - has been envisioned for natural language processing tasks in [59] leading to an improvement in performances on several NLP tasks with few training instances.

5. Synthesis of Appended Papers

In the following section, a synthesis of each of the appended papers is presented. Each paragraph will briefly present and discuss each paper, focusing on the specific goals and its main results. Figure 5 shows the mapping between the Open Issues identified in Section 3.1

[10] https://www.mitpressjournals.org/doi/pdf/10.1162/COLI$_{a0}$0239

along with the relative Research Areas identified in Section 4 and the following Research Outcomes:

1. "Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis" authored with Orsenigo C. and Vercellis C. (Politecnico di Milano - Italy)
2. "Unintended Bias in Misogyny Detection" authored with Nozza D. and E. (Università di Milano Bicocca - Italy)
3. "Temporal Word Embeddings for Narrative Understanding" authored with Antonucci A. and Kanjirangat V. (IDSIA - Istituto dalle Molle per l'Intelligenza Artificiale - Switzerland)

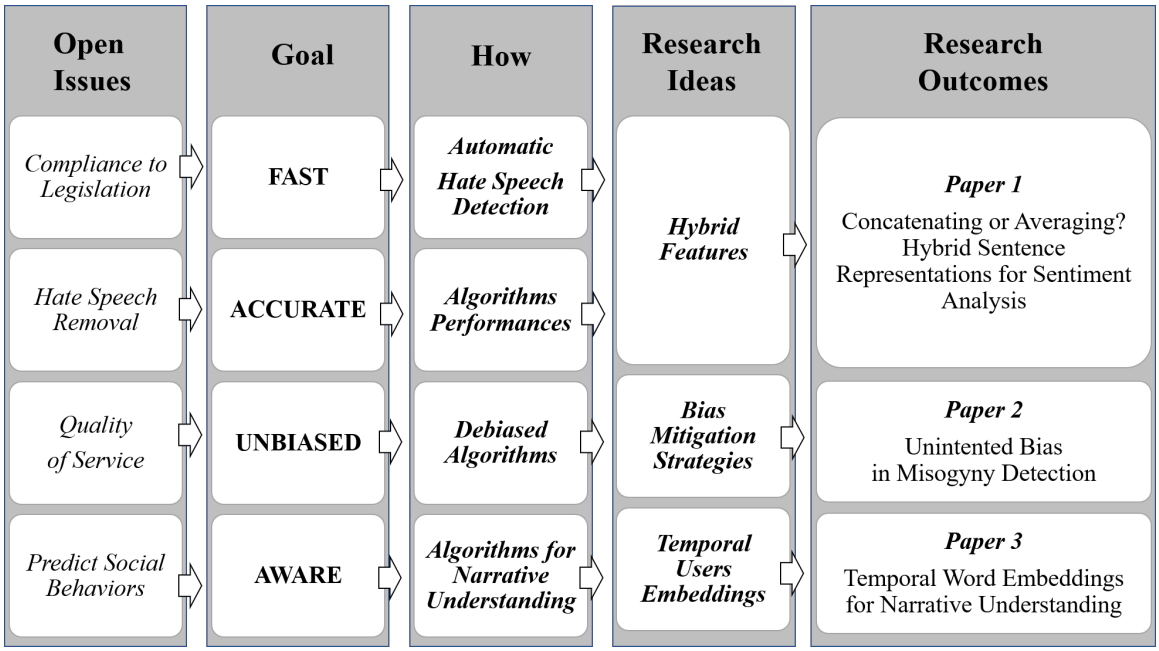

Fig. 5. Research Outcomes and Open Issues mapping process.

As Figure 5 illustrate, paper *"Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis"* aims at designing a new feature engineering technique in order to improve performances of text classification algorithms. So that to fulfill the needs of *faster* and *more accurate* text classification automatic systems. On the other hand, paper *"Unintended Bias for Misogyny Detection"* aims at fulfilling the need of *new bias mitigating techniques* in hate speech detection tasks in order to ensure *unbiased algorithms*. The paper indeed demonstrated the new strategies to be successful and methods have been applied in the specific case of misogyny detection. Finally, the last paper *"Temporal Word Embeddings for Narrative Understanding"* aims at investigating the idea that temporal distributed word representations can be used in order to identify character roles in narrative.

This is the first step into the direction of using those embeddings in social media in order to *identify social behaviors*. An important application of the identification and analysis of such character-centric narratives in social media could be the identification of victims and bullies in hate-speech dialogues.

5.1 PAPER 1 - Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis

As first research step, my study focused on *hybrid sentence representations of words in order to improve performances in text classification tasks and providing new hybrid variants which proved a stable increase in performances across different domains and datasets*. Main results of this first investigation in hybrid features representations are highlighted in the following of this paragraph. The ultimate goal of this study was to design brand new hybrid variants of sentence representations. My research investigated the effectiveness of combining the semantic features provided by Word2vec [53] with the lexical embedding generated by the well-established and more commonly used Bag of Word (BOW) model. The goal was to evaluate at what extent the Word2vec features complement and enrich the information comprised in the BOW representation and, specifically, to verify on an empirical basis whether the joint use of Word2vec and BOW in text classification for sentiment analysis leads to a sustainable performance improvement over the latter approach used alone. To this end, we designed and applied four hybrid sentence representations to convert textual data into numeric vectors, which benefit from both Word2vec and BOW information. These hybrid variants were compared against two baselines given by the classical BOW and Word2vec methods applied individually. Several tests in the context of sentiment classification were performed on five publicly available Amazon datasets, containing the users' opinions on products coming from different categories [60, 61]. The results of our experiments highlighted the usefulness of the novel hybrid representations across the different domains. In particular, the features obtained by concatenating the BOW model with the averaged form of Word2vec consistently outperformed the corresponding baselines.

New Hybrid Features Design. In my work I focused on features engineering rather than ensemble methods such in [62] since they present the drawback of high computational costs. In addition, I designed novel hybrid embedding variants resorting to different types of vectors concatenation which, at the best of my knowledge, haven't been tested yet. Unlike [63], I introduced the use of a presence-based approach as suggested in [49] and the simple average, which showed better performances, compared to the weighted approach, as reported in [62]. Specifically, by combining in all possible ways the BOW and Word2vec schemes, we derived the four hybrid representations of sentences:

$$avg_i \oplus bow01_i \tag{1}$$

$$avg_i \oplus bowTF_i \tag{2}$$

$$avgTF_i \oplus bow01_i \tag{3}$$

$$avgTF_i \oplus bowTF_i \tag{4}$$

Where vector $\vec{bow}01_i$ is the boolean vector representing sentence s_i whose generic element j takes the value 1 if and only if the word v_j appears in the sentence.

Where vector $\vec{bowTF_i}$ is vector representing sentence s_i containing at position j the TF-IDF value of word v_j, defined as the product between the term frequency, i.e. the number of occurrences of the word v_j in s_i, and the inverse term frequency, given by the logarithm of the ratio of the number of sentences divided by the number of sentences in the corpus containing v_j.

Where vector $\vec{avg}_i$ is a d-dimensional vector representing a sentence s_i whose k^{th} element is defined as:

$$\vec{avg}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} \vec{w}_{ijk} \quad k = 1, 2, \ldots, d, \tag{5}$$

where $\vec{w}_{ij}$ denotes the d-dimensional pre-trained word vector corresponding to the j^{th} term in sentence s_i and n_i is the number of words in sentence i.

Where the vector $\vec{avgTF}_i$ is a d-dimensional vector representing a sentence s_i whose element at position k is given by

$$\vec{avgTF}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} f_{ij} \vec{w}_{ijk} \quad k = 1, 2, \ldots, d, \tag{6}$$

where f_{ij} is the TF-IDF value of word j in sentence i. Notice that, for both strategies the vector representing each sentence has the same dimension d of the input word vectors. As a consequence, the original corpus is mapped into a numeric matrix of size $m \times d$ which can be fed to any machine learning classifier.

Experimental Framework and Main Results Overview. The following paragraph provides a brief report of main experimental settings and main results; for a full explanation of both methodologies and summary please refer to the full paper. To evaluate the effectiveness of the hybrid variants we performed several experiments on a publicly available[11] corpus of Amazon reviews [60, 61], referred to products from different categories such as Beauty, Video Games, Clothing, Health and Home as illustrated in Figure 6. Then for each dataset a preprocessing and a feature engineering step were performed before moving to the training and testing phases. Notice that Amazon datasets where chosen since their intrinsic diversity allowed to analyse the performances of the novel encoding schemes across different text domains. In order to provide a complete comparative study, computational tests

[11] http://jmcauley.ucsd.edu/data/amazon/

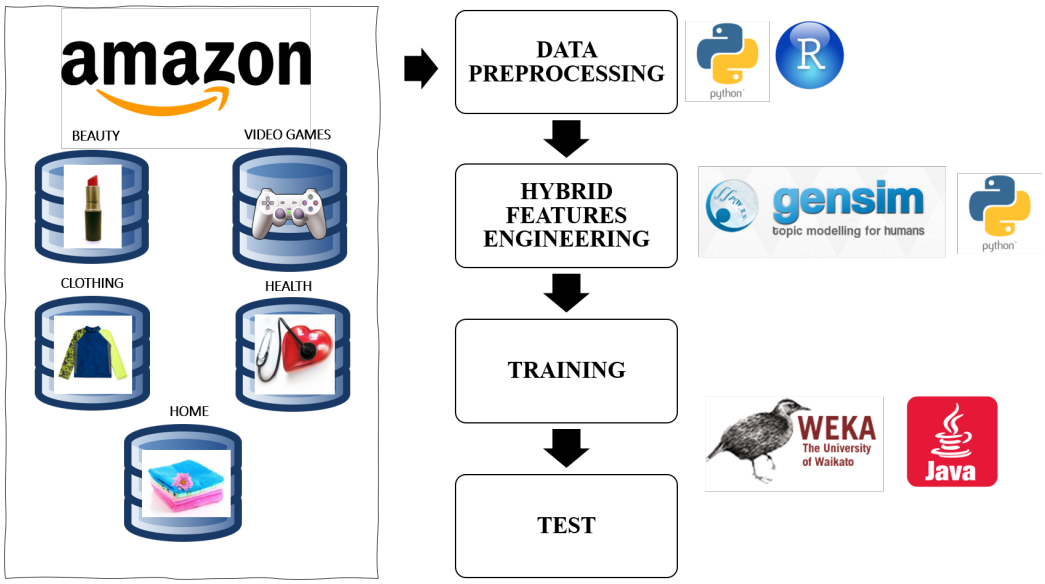

Fig. 6. Experimental Framework.

were performed on seven different sentence representations, composed by three baselines and four newly introduced hybrid variants. To discriminate between positive and negative reviews we resorted to the Logistic Regression classifier implemented in Weka [64]. Furthermore, given the high unbalance of the datasets in terms of class distribution, we selected as performance measure the F1-score on the minority (negative) class. The results of our experiments are shown in Table 3, which indicates the F1-score computed on the different test sets. The first main outcome is the notable performance exhibited by the hybrid encoding variants compared to the classical BOW and Word2vec representations. In particular, the avg_bow01 mapping obtained the highest accuracy across all the datasets. This empirical achievement emphasizes the benefits stemming from the joint use of BOW and Word2vec, suggesting that exploiting the information encoded in the two schemes by concatenating the corresponding sentence-level vectors outperforms the baseline approaches. Moreover, among the hybrid embeddings the one relying on the averaged Word2vec and on the presence-based BOW model consistently provided better results compared to the TF-IDF weighted alternatives. Finally, this evidence confirms that the weighting scheme based on TF-IDF, originally proposed to account for the importance of the words in a document within a corpus, is not always an effective choice in a sentiment analysis task.

Experiments confirmed the effectiveness of the hybrid mappings which showed notable performances in terms of prediction accuracy compared to the BOW and Word2vec approaches applied individually. Our empirical finding supports the results obtained in previous studies which conjectured on how the information provided by the well-established BOW scheme can be completed and enriched by the one contained in the more recently proposed Word2vec models. Research outcomes has been presented on November 22nd the *IDEAL 2018: 19TH International Conference on Intelligent Data Engineering and Au-*

Table 3. F1-scores on the test sets.

Representation	Beauty	Video Games	Clothing	Health	Home
*bow*01	*0.49*	*0.53*	0.51	*0.41*	*0.47*
avg	0.47	0.52	*0.52*	0.37	0.45
avgTF	0.38	0.40	0.42	0.29	0.35
$avg \oplus bow01$	**0.51**	**0.54**	**0.52**	**0.43**	**0.48**
$avg \oplus bowTF$	**0.51**	**0.54**	0.51	**0.43**	0.47
$avgTF \oplus bow01$	0.50	0.53	0.50	0.42	0.45
$avgTF \oplus bowTF$	0.50	0.53	0.50	0.42	0.45

tomated Learning, Session 10B: Natural Language Processing Computational Linguistics (Madrid - Spain) and published by Springer in the Lecture Notes in Artificial Intelligence - Scopus Indexed.

5.2 PAPER 2 - Unintended Bias in Misogyny Detection

This Section presents the context, the experimental framework and main results of the publication "Unintended Bias in Misogyny Detection" co-authored with Nozza, D. and Fersini E. of University Milano - Bicocca. Research outcomes will be presented on the 16 October 2019 at EEE WIC ACMInternational Conference on Web Intelligence (Thessaloniki, Greece) Session Web of People VII - Social vulnerabilities and tendencies.

Misogyny Detection. In the latest years, there was a growing interest in accelerating progress for women's empowerment and gender equality in our society. However, misogyny as a form of hate against them spread exponentially through the web and at very high-frequency rates, especially in online social media, where anonymity or pseudo-anonymity enables the possibility to afflict a target without being recognized or traced. This alarming phenomenon has triggered many studies related to the problem of abusive language recognition, and in particular for misogyny detection, both from computational linguistics and machine learning points of view. The state-of-the-art of automatic misogyny identification in online environments is still in its infancy. A preliminary exploratory analysis of misogynous language in online social media has been presented in [33], where the authors collected and manually labeled a set of tweets as positive, negative and neutral, providing some basic statistics about the usage of some candidate misogynistic keywords. A first contribution to the problem of automatic misogyny identification has been presented in [31], where the role of different linguistic features and machine learning models have been investigated. More recently, thanks to the Automatic Misogyny Identification (AMI) challenges organized at IberEval [65] and Evalita [66], many different approaches [67–72] have been proposed for addressing this problem. In this context, research works commonly focus on textual feature representation studying different linguistic characteristics, ranging from pragmatical, syntactical and lexical features to higher level features derived through

embedding techniques, or on the machine learning model, employing traditional or Deep Learning supervised models.

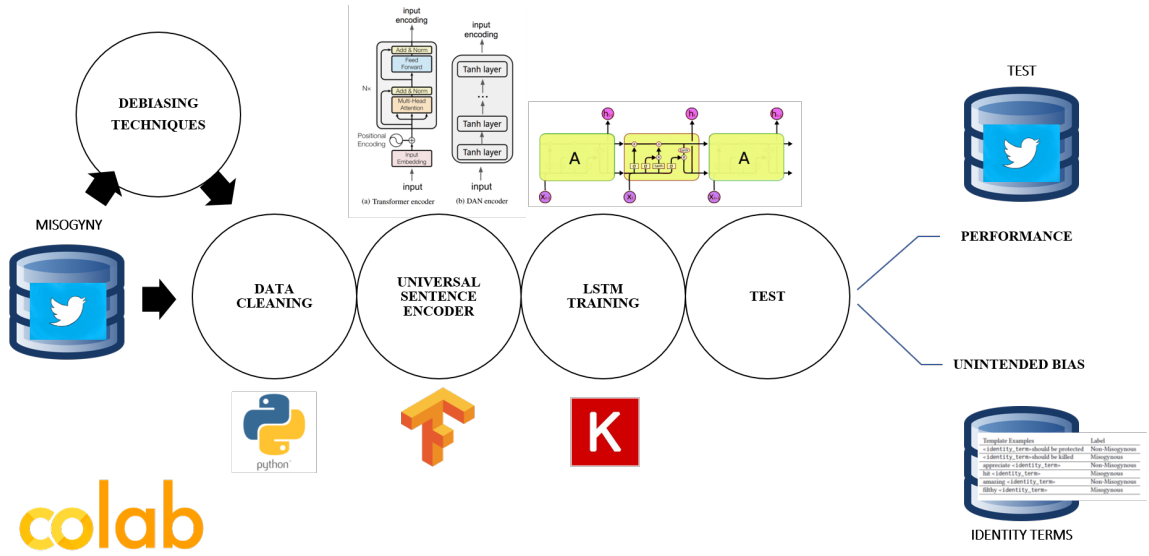

Fig. 7. Experimental Framework.

Unintended Bias in Hate Speech Detection algorithms. When training a text classification algorithms to perform abusive language classification, it is important to focus on a particular error induced by the training data, i.e. the bias introduced in the model by a set of *identity terms* that are frequently associated to the misogynous class. For example, the term *women*, if frequently used in misogynous messages, would lead most of the supervised classification models to associate an unreasonably high misogynous score to clearly non-misogynous text, such as "You are a woman". This behavior of recognition models is known as *unintended bias*. In particular, *"a model contains an unintended bias if it performs better for comments containing some particular identity terms than for comments containing others"* [73]. Tackling this error means being able to use those models in the real world. For the full investigation on the related work on this topic please refer to Section Annex A.

Experimental Framework and Main Results Overview. This paper was designed in order to provide a text classification algorithms for misogyny detection. The model proposed demonstrated to obtain the best classification performance in the state-of-the-art while we are also able to address the fairness of this model by measuring and mitigating its unintended bias. In particular, to address this challenge we first propose a novel synthetic template that can be also used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems. Additionally, we investigated different bias mitigation strategies, obtaining a debiased model that is less sensitive to identity terms as long as able to perform at the state of art of the best misogyny detection model in the litera-

ture on benchmark datasets. As *dataset*, we consider the state-of-the-art corpus for misogyny detection in the English language proposed for the Automatic Misogyny Identification shared task at the Evalita 2018 evaluation campaign [65]. The *identity terms* are chosen as those terms that can be used to refer to women, which may be unreasonably classified as misogynous with high scores. In order to define the list of identity terms, we take into consideration all the synonyms for "woman" by using a thesaurus[12]. The obtained list of synonymous has been then extended by including their plural form. Since some terms (e.g. *gentlewoman*) barely appear in the corpus, we decided to remove the ones with a frequency lower than 3. Since unintended bias of identity terms cannot be measured on the original test set due to class imbalance and highly different identity term contexts, *synthetic test sets* has been generated on purpose and building on the previous work [73], we manually created a balanced synthetic dataset of misogynous and non-misogynous contents. As *classification models*, we first built a machine learning model on the state-of-the-art misogyny corpus proposed in [65]. We first encoded the English sentences using the *Universal Sentence Encoder* introduced in [74] built using a transformer architecture [75] and available online[13]. Once constructed the sentence embeddings, we used them as input to a single-layer neural network architecture and trained what we called the USE_T model. Then we created *four debiased versions* of our USE_T model in order to mitigate its bias. The first one consists of mitigating the class imbalance of the identity terms which have the most imbalanced class distributions and it is called *Debiased*. Moreover, we also build the *Debiased_length* model, which is trained on a debiased set where the class balance is obtained also considering tweet length ranges. In order to confirm the benefits of the described bias mitigation procedure instead of a simple data augmentation process, we investigate the addition of randomly sampled data from the external corpus, so we obtained two bias mitigated models called *Random* and *Random_length* model. *As performance metric*, we adopted the AUC (area under the curve) measure to evaluate the classification performance of the misogyny detection model on the test set and on the synthetic dataset. *As measure for the unintended bias*, we computed the metrics introduced in recent state-of-the-art works [73, 76] to measure the extent of unintended bias in the model.

Summarizing the experiments results, this paper demonstrated that the bias mitigation strategies have significantly decreased the false positive and false negative rates for each identity term and *consequently reduced the unintended bias* by providing more similar values across terms. The debiased model showed a stable improvement in separability of positive and negative examples within each subgroup, if compared with the reference model subgroups. Additionally, the paper first propose a novel synthetic template set that can be used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems.

[12] www.thesaurus.com

[13] https://tfhub.dev/google/universal-sentence-encoder-large/3

5.3 PAPER 3 - Temporal Word Embeddings for Narrative Understanding

In this paper, we studied *temporal word embeddings* as a possible tool for effective narrative understanding. This work is first step towards the use of temporal embeddings to understand the social behaviors in the context of other narratives e.g. narratives from social media. An important application of the identification and analysis of such narratives in social media could be the identification of victims and bullies in hate-speech dialogues. In this work, first we advocate and demonstrate the efficiency of using temporal word embeddings for narrative understanding. Second, we provided a new data set of temporal word analogies and we also tested a new variant of a recently proposed temporal embedding approach. Results showed a good accuracy when solving temporal character analogies across time. This supports that research idea that these embeddings can properly understand the semantic role of each character. We also provided a visualization of the temporal embeddings to trace the evolution over time of characters in a story plot.

Automatic identification of characters in a Story: the Idea. In this paper we aim at automatically identifying characters roles and their evolution over time. A *character role* describes what function a character serves in the story. The *character evolution* is the idea in writing that a character can ideally change from the beginning of a work to the last sentence, e.g., from the villain to the hero. To validate these techniques, we use J.K. Rowling's Harry Potter books as a benchmark providing a consistent amount of text, with a story spread over multiple books with recurrent characters and varying relations among them.

Our claim is that if we construct Temporal Word Embeddings (TWEs) (also known as dynamic word embeddings [77] or diachronic word embeddings [78–80]) for each character in different time periods (e.g., for each character in each book of a book series) they can be used to represent the role and the evolution of a character along a story plot. TWEs make it possible to find distinct words that share a similar meaning in different periods of time by retrieving temporal embeddings that occupy similar regions in the vector spaces that correspond to distinct time periods. Consequently, our hypothesis is that characters having the same role, they are assumed to be closer, according to some metric measure in the embedding space, to similar characters that producing a distribution in the vector space so that, e.g., villains should be clustered in a different area from the area in which story heroes are placed. On the other hand, we also claim that by building a sequence of temporal embeddings of a character over consecutive time intervals, one can track the character evolution (semantic shift) occurred in the character role.

Temporal Word Embeddings models explained. In this paragraph we report an informal descriptions of the TWEs models implemented in our paper in order to let the reader gain the main intuition of the designed approaches without the barrier of a specialized terminology. In Annex A, the full paper with a more formal definition of the TWEs models is provided. In this work, we build on the specific method introduced in [81]. This method assumes that a word, e.g., Clinton appears during some temporal periods in the contexts of words that are related to his position, e.g., president, that conversely doesn't change its

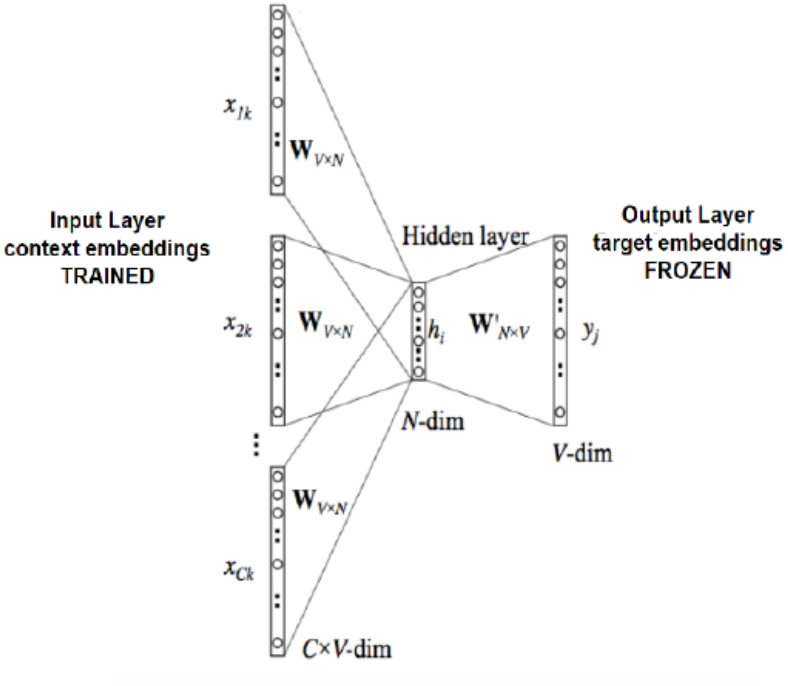

Fig. 8. Training input vectors with frozen output vectors.

meaning through time. This assumption allows to heuristically consider the context matrix of a CBOW [52] training system (Figure 8) as static, i.e., to freeze the output weight matrix during training, while allowing the word embedding input weight matrices, to change on the basis of co-occurrence frequencies that are specific to a given temporal interval (Figure 8). After training, model returns the context embeddings, that we are going to consider as a TWE. This is achieved by a two-fold training procedure. First a static word embedding is trained, with random initialization, using the entire vocabulary and ignoring temporal slices. This initialization has been proved to force alignment and make it possible to compare vectors from embeddings associated to different time slices. In this paper we propose a different initialization scheme for such a training architecture. In the particular case of the characters of a narrative text, as basically each character might change their semantic position over time, we designed a better initialization strategy. It consists in using the matrix resulting from the previous training step *t-1* as the initialization of matrix at time *t* for each time step. We call this procedure dynamic initialization, while the original procedure proposed in [81] is called here static initialization. A graphical summary of the architecture together with the two initialization strategies is depicted in Figure 9.

Experimental Framework and Main Results Overview. For our experiments we considered as a *corpus* the six books from the Harry Potter's series. Since the training process of a TWEs relies on diachronic text corpora, we need to decompose our corpus into temporal slices which are usually, temporal intervals set accordingly to the granularity of time

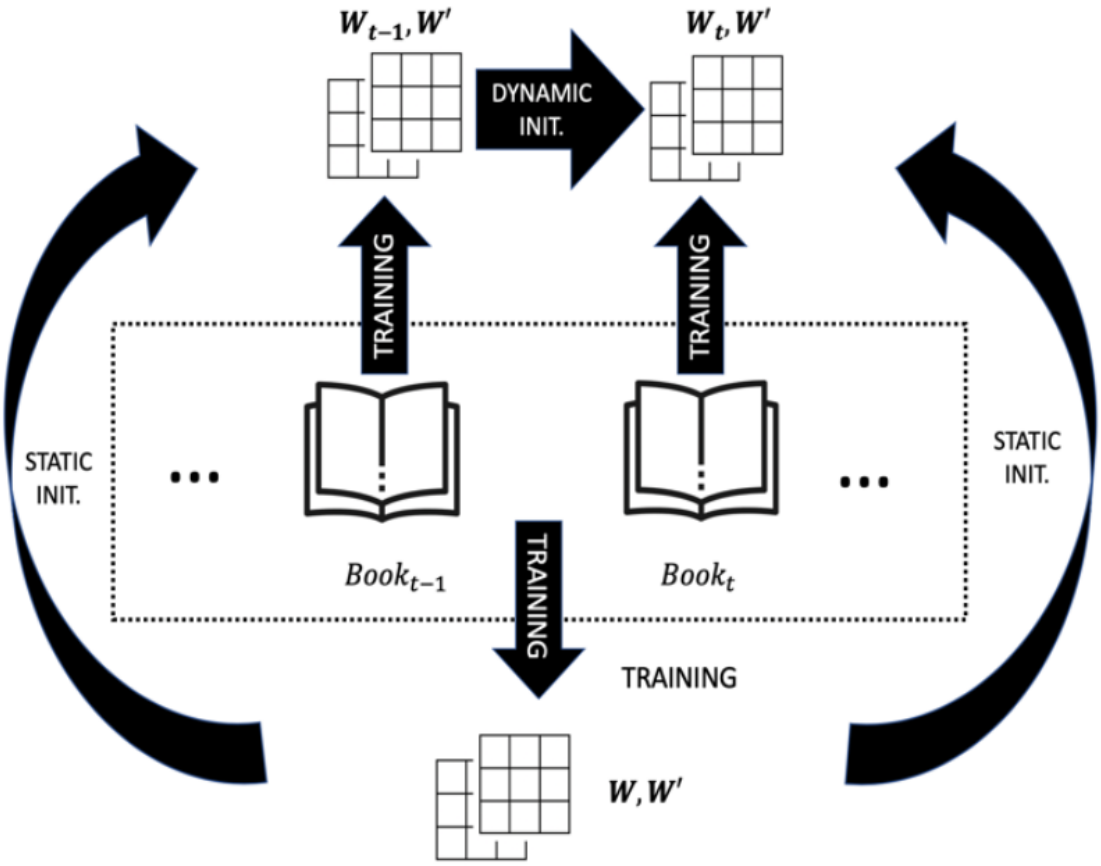

Fig. 9. Temporal context embeddings architecture with both static and dynamic initialization.

spans we want to cover. Since we are trying to trace major changes in characters behaviors and role, we decided to keep the granularity of time spans low and consequently we set our *time unit* (the granularity of the temporal dimension) to the number of books. As a result of this choice, after the training every word will have six representations, one per each time unit (per each book). In order to build the Temporal Word Analogies test set, we asked ten "experts", i.e., people who carefully and repeatedly read the six books, to answer a survey. They were asked to answer twelve questions about Harry Potter's characters across the first six books. This approach made it possible to trace a series of 150 characters analogies over time. We then applied Named entity recognition tools form the API of the NLP Stanford to retrieve the characters from the corpus and finally we trained both models (see previous section) on the entire Harry Potter corpus (six books) to build the static embeddings and then we *trained* separately the temporal embeddings according the two different approaches discussed in the previous paragraph. After some hyperparameters tuning procedures, we fixed 200 for the word embeddings dimensionality, we specified a window size equal to two, and twenty epochs for the static training. Temporal embeddings were trained for five epochs each, when replicating the original approach, and gradually

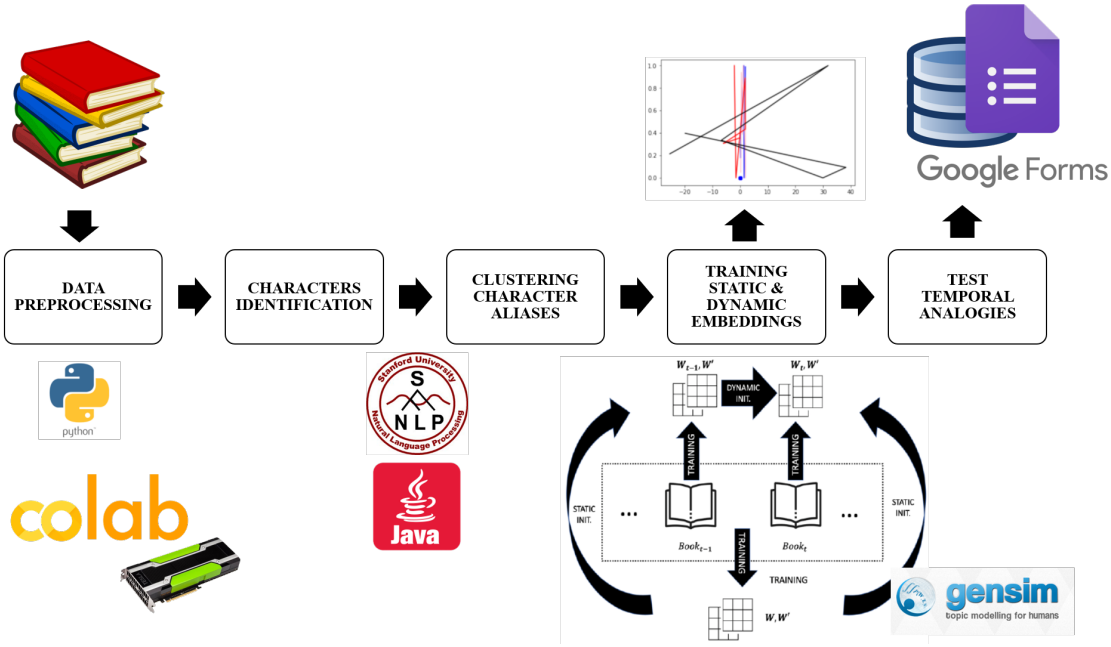

Fig. 10. Experimental Framework.

scaled from ten to one when dealing with the second approach. The resulting visualization plots and accuracy metrics clearly demonstrated that our approach is well suited for this research problem. Both visualization techniques and the testing on the temporal word analogies show that these embeddings can properly understand the semantic role of each character. As a future work, we would like to use those embeddings for analogous tasks such as understanding narratives from social media.

6. Conclusions and Future Developments

The aim of this thesis was to prove that new artificial intelligence algorithms could be helpful in providing research stakeholders with a *faster, accurate, unbiased and aware* automatic system to detect hateful comments. Newly designed methods have been introduced namely: *hybrid sentence representations*, *unintended bias mitigation techniques for misogyny detection* and *temporal word embeddings for narrative understanding*. Research outcomes have been demonstrating promising results and demonstrated to support the research intuitions behind this PhD work. The first publication addressed the need of *faster* and *accurate* algorithms by designing novel hybrid sentence representations able to exploit the information provided by two different strategies coming from the two different worlds of machine and deep learning. Experiments confirmed the effectiveness of the hybrid mappings which showed notable performances in terms of prediction accuracy. But more importantly, from an academic point of view, our empirical finding supports the results obtained in previous studies which conjectured on how the information provided by the two representations are complementary. The second research effort answered to the need of new strategies for measuring and mitigating *unintended bias* in hate speech detection tasks. Results of this work confirmed the ability of the bias mitigation treatment implemented to reduce the unintended bias of the proposed misogyny detection model. From a practitioners point of view, this work represent a further step towards the design of models that are robust to training biases and that consequently can be used in real world applications because of their capability to ensure a fair and unbiased service with respect to every group of people without discrimination. The third research work aimed at satisfying the stakeholders' need of a better understanding of *social behaviors* behind online discussions by providing a new training approach for temporal word embeddings. These were compared against human annotators experts and they showed promising results as tool for narrative understanding. This supports the intuition that these embeddings can properly understand the semantic role of each user involved in discourse and consequently the role of victims and agressors in hate speech online discourses. *This PhD thesis provided a further step towards the construction of safer and unbiased virtual spaces for users discussions*. As a future work, an important development that can be built especially on the last research paper outcomes, is the identification and analysis of narratives in social media in order to be able to identify e.g. victims and bullies in hate speech dialogues. More in general, understanding social media content is a process that more and more is demanding to *anticipate and predict serious hate events* such as cyberbullying or off-line violence. Addressing such expectations is going to be crucial in the next decade when public authorities will be pushing forward their regulatory effort by asking social media companies for more and more awareness on the content published on their platforms. On the other hand, as further future development, this work can be considered a first step in the direction of investigating the interplay between Sentiment Analysis and Hate Speech Detection. As second step in the direction of further leveraging the interplay between these tasks, we envision the use of Transfer Learning between Sentiment Analysis and Hate Speech Detection.

References

[1] Mäntylä MV, Graziotin D, Kuutila M (2018) The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review* 27:16–32. https://doi.org/10.1016/J.COSREV.2017.10.002

[2] Piryani R, Madhavi D, Singh V (2017) Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management* 53(1):122–150. https://doi.org/10.1016/J.IPM.2016.07.001

[3] Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology* 60(11):2169–2188. https://doi.org/10.1002/asi.21149

[4] Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2011) Election Forecasts With Twitter. *Social Science Computer Review* 29(4):402–418. https://doi.org/10.1177/0894439310386557

[5] Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. *Journal of Computational Science* 2(1):1–8. https://doi.org/10.1016/J.JOCS.2010.12.007

[6] Nockleby JT (2000) Hate Speech. *In Encyclopedia of the American Constitution (2nd ed, edited by Leonard W Levy, Kenneth L Karst et al, New York: Macmillan, 2000)* pp. 1277–.

[7] Alex Hern (2016) Facebook, YouTube, Twitter and Microsoft sign EU hate speech code — Technology — The Guardian. Available at https://www.theguardian.com/technology/2016/may/31/facebook-youtube-twitter-microsoft-eu-hate-speech-code.

[8] Fortuna P, Nunes S (2018) A survey on automatic detection of hate speech in text. *ACM Computing Surveys* 51(4). https://doi.org/10.1145/3232676. Available at https://www.scopus.com/inward/record.uri?eid=2-s2.0-85053245173{&}doi=10.1145{%}2F3232676{&}partnerID=40{&}md5=e92c863dbc8e557ce6410de00d4c4f3c

[9] Schmidt A, Wiegand M (2017) A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media* (2012):1–10.

[10] American Association for Artificial Intelligence CR, IAAI Conference (8th : 1996 : Portland O (1996) *Proceedings of the Thirteenth National Conference on Artificial Intelligence and the Eighth Innovative Applications of Artificial Intelligence conference* (AAAI Press), . Available at https://dl.acm.org/citation.cfm?id=1867616.

[11] Razavi AH, Inkpen D, Uritsky S, Matwin S (2010) Offensive language detection using multi-level classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer-Verlag), Vol. 6085 LNAI, pp 16–27. https://doi.org/10.1007/978-3-642-13059-5_5. Available at http://link.springer.com/10.1007/978-3-642-13059-5{_}5

[12] Sood SO, Antin J, Churchill EF (2012) Profanity use in online communities. *Conference on Human Factors in Computing Systems - Proceedings* (ACM Press, New York, New York, USA), , pp 1481–1490. https://doi.org/10.1145/2207676.2208610.

Available at http://dl.acm.org/citation.cfm?doid=2207676.2208610

[13] Xiang G, Fan B, Wang L, Hong J, Rose C (2012) Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12* (ACM Press, New York, New York, USA), , p 1980. https://doi.org/10.1145/2396761.2398556. Available at http://dl.acm.org/citation.cfm?doid=2396761.2398556

[14] Burnap P, Williams ML (2014) Hate speech, machine classification and statistical modelling of information flows on Twitter: interpretation and communication for policy decision making Available at http://orca.cf.ac.uk/65227/.

[15] Warner W, Hirschberg J (2012) Detecting Hate Speech on the World Wide Web :19–26Available at http://delivery.acm.org/10.1145/2400000/2390377/p19-warner.pdf?ip=131.175.11.27{&}id=2390377{&}acc=OPEN{&}key=296E2ED678667973.7773E6D96819F65E.4D4702B0C3E38B35.6D218144511F3437{&}CFID=1002655736{&}CFTOKEN=27087067{&}{_}{_}acm{_}{_}=1509984931{_}10ebc691453adcab7a537196b7f9f2c.

[16] Burnap P, Williams M (2015) Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy and Internet* 7(2). https://doi.org/10.1002/poi3.85

[17] Silva L, Mondal M, Correa D, Benevenuto F, Weber I (2016) Analyzing the targets of hate in online social media. *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*, , pp 687–690.

[18] Kwok I, Wang Y (2013) Locate the hate: Detecting tweets against blacks. *Proceedings of the 27th AAAI Conference on Artificial Intelligence, AAAI 2013*, , pp 1621–1622.

[19] Ross B, Rist M, Carbonell G, Cabrera B, Kurowsky N, Wojatzki M (2016) Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *Bochumer Linguistische Arbeitsberichte 17 (BLA 17), Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication* :6–9https://doi.org/10.17185/duepublico/42132. arXiv:1701.08118v1

[20] Xu JM, Jun KS, Zhu X, Bellmore A (2012) Learning from bullying traces in social media. *NAACL HLT 2012 - 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, , pp 656–666. Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.374.1862.

[21] Hosseinmardi H, Mattson SA, Rafiq RI, Han R, Lv Q, Mishra S (2015) Analyzing labeled cyberbullying incidents on the instagram social network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 9471, pp 49–66. https://doi.org/10.1007/978-3-319-27433-1_4. 1503.03909 Available at http://arxiv.org/abs/1503.03909

[22] Thompson N (2016) *Anti-discriminatory practice: Equality, diversity and social justice* (Macmillan International Higher Education), .

[23] Jigsaw (2017) Perspective API. Available at https://www.perspectiveapi.com/{#}

/home.

[24] Guermazi R, Hammami M, Hamadou AB (2007) Using a semi-automatic keyword dictionary for improving violent web site filtering. *Proceedings - International Conference on Signal Image Technologies and Internet Based Systems, SITIS 2007* (IEEE), , pp 337–344. https://doi.org/10.1109/SITIS.2007.137. Available at http://ieeexplore.ieee.org/document/4618794/

[25] McNamee LG, Peterson BL, Peña J (2010) A Call to Educate, Participate, Invoke and Indict: Understanding the Communication of Online Hate Groups. *Communication Monographs* 77(2):257–280. https://doi.org/10.1080/03637751003758227. Available at http://www.tandfonline.com/doi/abs/10.1080/03637751003758227

[26] Prentice S, Taylor PJ, Rayson P, Hoskins A, O'Loughlin B (2011) Analyzing the semantic content and persuasive composition of extremist media: A case study of texts produced during the Gaza conflict. *Information Systems Frontiers* 13(1):61–73. https://doi.org/10.1007/s10796-010-9272-y. Available at http://link.springer.com/10.1007/s10796-010-9272-y

[27] Agarwal S, Sureka A (2015) Using KNN and SVM Based One-Class Classifier for Detecting Online Radicalization on Twitter (Springer, Cham), , pp 431–442. https://doi.org/10.1007/978-3-319-14977-6_47. Available at http://link.springer.com/10.1007/978-3-319-14977-6{_}47

[28] Douglass S, Mirpuri S, English D, Yip T (2016) "They were just making jokes": Ethnic/racial teasing and discrimination among adolescents. *Cultural diversity & ethnic minority psychology* 22(1):69–82. https://doi.org/10.1037/cdp0000041. Available at http://www.ncbi.nlm.nih.gov/pubmed/26009942http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4659767

[29] Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification, , .

[30] Borkan D, Sorensen J, Dixon L, Vasserman L, Thain N (2019) Nuanced metrics for measuring unintended bias with real data for text classification, , .

[31] Anzovino M, Fersini E, Rosso P (2018) Automatic identification and classification of misogynistic language on twitter. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Springer, Cham), Vol. 10859 LNCS, pp 57–64. https://doi.org/10.1007/978-3-319-91947-8_6. Available at http://link.springer.com/10.1007/978-3-319-91947-8{_}6

[32] Sanguinetti M, Poletto F, Bosco C, Patti V, Stranisci M (2018) An Italian twitter corpus of hate speech against immigrants. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)* (European Languages Resources Association (ELRA), Miyazaki, Japan), , . Available at https://www.aclweb.org/anthology/L18-1443.

[33] Hewitt S, Tiropanis T, Bokhove C (2016) The problem of identifying misogynist language on Twitter (and other online social spaces). *Proceedings of the 8th ACM Conference on Web Science - WebSci '16* (ACM Press, New York, New York, USA),

, pp 333–335. https://doi.org/10.1145/2908131.2908183. Available at http://dl.acm.org/citation.cfm?doid=2908131.2908183

[34] Nakagawa T, Inui K, Kurohashi S, Technology C (2010) Dependency Tree-based Sentiment Classification using CRFs with Hidden Variables (June):786–794.

[35] Silva J, Coheur L, Mendes AC, Wichert A (2011) From symbolic to sub-symbolic information in question classification. *Artificial Intelligence Review* 35(2):137–154. https://doi.org/10.1007/s10462-010-9188-4

[36] Socher R, Pennington J, Huang EH, Ng AY, Manning CD (2011) Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions. *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (ii):151–161. https://doi.org/10.1.1.224.9432. Available at http://dl.acm.org/citation.cfm?id=2145450

[37] Wang S, Manning C (2012) Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics* (July):90–94.

[38] Hermann KM, Blunsom P (2013) The Role of Syntax in Vector Space Models of Compositional Semantics. *Acl (1)* :894–904Available at http://www.aclweb.org/anthology/P13-1088%5Cnhttp://www.karlmoritz.com/_media/hermannblunsom_acl2013.pdf.

[39] Socher R, Perelygin A, Wu J (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Emnlp* :1631–1642https://doi.org/10.1371/journal.pone.0073791. Available at http://nlp.stanford.edu/~socherr/EMNLP2013_RNTN.pdf%5Cnhttp://www.aclweb.org/anthology/D13-1170%5Cnhttp://aclweb.org/supplementals/D/D13/D13-1170.Attachment.pdf%5Cnhttp://oldsite.aclweb.org/anthology-new/D/D13/D13-1170.pdf

[40] Wang SI, Manning CD (2013) Fast dropout training. *Proceedings of the 30th International Conference on Machine Learning* 28:118–126. Available at http://machinelearning.wustl.edu/mlpapers/papers/wang13a.

[41] Li S, Zhang H, Xu W, Guo J (2014) Chinese Text Sentiment Analysis Based on Combination Model (10):1–7.

[42] Cardie C Deep Recursive Neural Networks for Compositionality in Language :1–9.

[43] Kalchbrenner N, Grefenstette E, Blunsom P (2014) A Convolutional Neural Network for Modelling Sentences. *In Proceedings of ACL* :655–665.

[44] Kim Y (2014) Convolutional Neural Networks for Sentence Classification :1746–1751https://doi.org/10.3115/v1/D14-1181. Available at http://arxiv.org/abs/1408.5882

[45] Le QV, Mikolov T (2014) Distributed Representations of Sentences and Documents 32. https://doi.org/10.1145/2740908.2742760. Available at http://arxiv.org/abs/1405.4053

[46] Yang B, Cardie C (2014) Context-aware Learning for Sentence-level Sentiment Analysis with Posterior Regularization. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2008):325–335.

Available at http://www.aclweb.org/anthology/P14-1031.

[47] Tai KS, Socher R, Manning CD (2015) Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks :1556–1566https://doi.org/10.1515/popets-2015-0023. Available at http://arxiv.org/abs/1503.00075

[48] Kumar A, Irsoy O, Su J, Bradbury J, English R, Pierce B, Ondruska P, Gulrajani I, Socher R (2016) Ask me anything: Dynamic memory networks for natural language processing. *Icml* 48.

[49] Pang B, Lee L (2008) *Opinion Mining and Sentiment Analysis*. Vol. 2 (Now Publishers, Inc.), . https://doi.org/10.1561/1500000011. 0112017

[50] Li X, Roth D (2002) Learning question classifiers. *Proceedings of the 19th International Conference on Computational Linguistics - Volume 1* COLING '02 (Association for Computational Linguistics, Stroudsburg, PA, USA), , pp 1–7. https://doi.org/10.3115/1072228.1072378. Available at https://doi.org/10.3115/1072228.1072378

[51] Wiebe J, Wilson T, Cardie C (2005) Annotating Expressions of Opinions and Emotions in Language. *Language Resources and Evaluation* 39(2-3):165–210. https://doi.org/10.1007/s10579-005-7880-9. Available at http://link.springer.com/10.1007/s10579-005-7880-9

[52] Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781. 1301.3781 Available at http://arxiv.org/abs/1301.3781.

[53] Socher, Richard and Perelygin, Alex and Wu, Jean and Chuang, Jason and Manning, Christopher D and Ng, Andrew and Potts C (2013) Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of EMNLP 2013* (Association for Computational Linguistics (ACL)), , pp 1631—-1642.

[54] Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735. Available at https://doi.org/10.1162/neco.1997.9.8.1735

[55] Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, , . Available at http://arxiv.org/abs/1409.0473.

[56] Ma Y, Peng H, Cambria E (2018) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *Thirty-Second AAAI Conference on Artificial Intelligence*, , .

[57] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, , pp 5998–6008. Available at http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[58] Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguis-*

tics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), , pp 4171–4186. Available at https://www.aclweb.org/anthology/N19-1423/.

[59] Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging nlp applications. *arXiv preprint arXiv:190602829* .

[60] McAuley J, Targett C, Shi Q, van den Hengel A (2015) Image-Based Recommendations on Styles and Substitutes. *Proceedings of SIGIR '15* (ACM Press, New York, New York, USA), , pp 43–52. https://doi.org/10.1145/2766462.2767755

[61] McAuley J, Pandey R, Leskovec J (2015) Inferring Networks of Substitutable and Complementary Products. *Proceedings of the ACM SIGKDD'15* (ACM Press, New York, New York, USA), , pp 785–794. https://doi.org/10.1145/2783258.2783381

[62] Enríquez F, Troyano JA, López-Solaz T (2016) An approach to the use of word embeddings in an opinion classification task. *Expert Systems with Applications* 66:1–6. https://doi.org/10.1016/j.eswa.2016.09.005

[63] Lilleberg J, Zhu Y, Zhang Y (2015) Support vector machines and Word2vec for text classification with semantic features. *Proceedings of 2015 IEEE 14th IICCI*CC* (IEEE), , pp 136–140. https://doi.org/10.1109/ICCI-CC.2015.7259377

[64] Witten IH, Frank E, Hall MA, Pal CJ (2016) *Data Mining: Practical Machine Learning Tools and Techniques* (Elsevier Inc.), .

[65] Fersini E, Nozza D, Rosso P (2018) Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* (CEUR-WS.org), , .

[66] Fersini E, Rosso P, Anzovino M (2018) Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (CEUR-WS.org), , .

[67] Pamungkas EW, Cignarella AT, Basile V, Patti V (2018) 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (CEUR-WS.org), Vol. 2150, pp 234–241.

[68] Frenda S, Ghanem B, Montes-y Gómez M (2018) Exploration of Misogyny in Spanish and English tweets. *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)* (CEUR-WS.org), Vol. 2150, pp 260–267.

[69] Goenaga I, Atutxa A, Gojenola K, Casillas A, de Ilarraza AD, Ezeiza N, Oronoz M, Pérez A, Perez-de-Viñaspre O (2018) Automatic Misogyny Identification Using Neural Networks. *Proceedings of the Third Workshop on Evaluation of Human Language*

Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018) (CEUR-WS.org), , .

[70] Pamungkas EW, Cignarella AT, Basile V, Patti V (2018) Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* (CEUR.org, Turin, Italy), , .

[71] Basile A, Rubagotti C (2018) CrotoneMilano for AMI at Evalita2018. A Performant, Cross-lingual Misogyny Detection System. *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)* (CEUR.org, Turin, Italy), , .

[72] Saha P, Mathew B, Goyal P, Mukherjee A (2018) Hateminers: Detecting hate speech against women. *arXiv preprint arXiv:181206700* .

[73] Dixon L, Li J, Sorensen J, Thain N, Vasserman L (2018) Measuring and mitigating unintended bias in text classification. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (ACM), , pp 67–73.

[74] Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, Constant N, Guajardo-Cespedes M, Yuan S, Tar C, et al. (2018) Universal sentence encoder. *arXiv preprint arXiv:180311175* .

[75] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is All you Need. *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, , pp 6000–6010. Available at http://papers.nips.cc/paper/7181-attention-is-all-you-need.

[76] Borkan D, Dixon L, Sorensen J, Thain N, Vasserman L (2019) Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. *arXiv preprint arXiv:190304561* .

[77] Yao Z, Sun Y, Ding W, Rao N, Xiong H (2018) Dynamic word embeddings for evolving semantic discovery. *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, Vol. 2018-Febua, pp 673–681. https://doi.org/10.1145/3159652.3159703. 1703.00607 Available at http://arxiv.org/abs/1703.00607http://dx.doi.org/10.1145/3159652.3159703

[78] Kutuzov A, Øvrelid L, Szymanski T, Velldal E (2018) Diachronic word embeddings and semantic shifts: a survey 1806.03537 Available at https://books.google.com/ngramshttp://arxiv.org/abs/1806.03537.

[79] Szymanski T Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings :448–453https://doi.org/10.18653/v1/P17-2071. Available at https://doi.org/10.18653/v1/P17-2071

[80] Hamilton WL, Leskovec J, Jurafsky D (2016) Diachronic word embeddings reveal statistical laws of semantic change. *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, Vol. 3, pp 1489–1501. Available at

http://nlp.stanford.edu/projects/histwords.

[81] Di Carlo V, Bianchi F, Palmonari M (2019) Training Temporal Word Embeddings with a Compass. *Proceedings of the AAAI Conference on Artificial Intelligence* 33:6326–6334. https://doi.org/10.1609/aaai.v33i01.33016326. 1906.02376 Available at http://arxiv.org/abs/1906.02376http://www.aaai.org/ojs/index.php/AAAI/article/view/4594

Annex A: Appended Papers

An annex that presents the final manuscripts for each of the papers.

Concatenating or Averaging? Hybrid Sentences Representations for Sentiment Analysis

Carlotta Orsenigo, Carlo Vercellis, and Claudia Volpetti(✉)

Department of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4b, 20156 Milan, Italy
{carlotta.orsenigo,carlo.vercellis,claudia.volpetti}@polimi.it

Abstract. Performances in sentiment analysis - the crucial task of automatically classifying the huge amount of users' opinions generated online - heavily rely on the representation used to transform words or sentences into numbers. In the field of machine learning for sentiment analysis the most common embedding is the bag of words (BOW) model, which works well in practice but which is essentially a lexical conversion. Another well-known method is the Word2vec approach which, instead, attempts to capture the meaning of the terms. Given the complementarity of the information encoded in the two models, the knowledge offered by Word2vec can be helpful to enrich the information comprised in the BOW scheme. Based on this assumption we designed and tested four hybrid sentence representations which combine the two former approaches. Experiments performed on publicly available datasets confirm the effectiveness of the hybrid embeddings which led to a stable increase in the performances across different sentiment analysis domains.

Keywords: Text classification · Sentiment analysis
Machine learning · Word vectors · Word2vec · Bag of words
Hybrid sentence representation

1 Introduction

With the rapid growth of opinionated texts produced daily by online users, the ability of classifying opinions has become imperative to understand political orientations [18], brand perception [5] or even to forecast the impact of news on financial markets [2]. Consequently, several research efforts in the machine learning domain have been recently devoted to design accurate text classification algorithms for sentiment analysis, in order to automatically assign a sentiment polarity to user-generated comments [9,16].

To apply machine learning, sentences must be converted into a numeric format through a vector-based representation. Vectors can be derived directly from the raw text by means of several strategies which reflect different lexical, syntactical or semantic properties of the documents. Choosing a specific text

H. Yin et al. (Eds.): IDEAL 2018, LNCS 11314, pp. 1–9, 2018.
https://doi.org/10.1007/978-3-030-03493-1_59

representation is crucial since it determines the information provided in input to the classifier and, therefore, considerably affects its performances [7].

For text classification the most common feature extraction method is the bag of words (BOW) model, in which each document is described in terms of a vocabulary of words built on a given training corpus. In the resulting dataset the numeric features indicate the presence or, alternatively, the frequency of each term in the document, where the former strategy has been shown to perform better compared to the latter for the purpose of sentiment classification [15]. Notice that, the bag of words approach discards any positional information about the terms in the sentences and relies only on the presence of a word or on how frequently the term occurs. Despite the effectiveness observed in practice, it is indeed an exclusively lexical transformation in which the word order is disrupted and syntactic structures are broken.

As an alternative to the BOW sparse vectors, some distributed representations of words as real-valued vectors, called word vectors, have been proposed to capture their semantic aspects [1]. Among these, the Word2vec model developed by [12] generates numeric vectors using neural networks, with the aim of grasping the meaning of each word considering its relations with other terms in the same context. [12] proposes two methods to learn these vectors from text data. The first is the Skip-Gram model which learns the word vectors by training a neural network to predict the surrounding context words given a central word. The second, called continuous bag of words, learns these vectors by predicting the central word given its context of surrounding terms in a fixed window. In both cases, the size of the window of neighboring terms is a parameter of the model. Since the construction of such representations is unsupervised and generally involves huge corpora of documents, such as Wikipedia [3] or Google News [12], a common strategy is to use pre-trained vectors instead of training them every time from scratch. The Word2vec representation bears some advantages over the BOW counterpart, since terms which are near in meaning are close in the word embedding space and, somewhat surprisingly, other semantic relationships such as gender-inflections or geographical connections can be recovered using algebraic operations between vectors [13,14]. Therefore, if the purpose is to extract semantic features able to bring some extra information related to the similarities among words, one may effectively resort to the Word2vec models which widely proved their ability to encode such extra knowledge.

The aim of the present study is to investigate the effectiveness of combining the semantic features provided by Word2vec with the lexical embedding generated by the well-established and more commonly used BOW model. In particular, our goal is to evaluate at what extent the Word2vec features complement and enrich the information comprised in the BOW representation and, specifically, to verify on an empirical basis whether the joint use of Word2vec and BOW in text classification for sentiment analysis leads to a sustainable performance improvement over the latter approach used alone. To this end, we designed and applied four hybrid sentence representations to convert textual data into numeric vectors, which benefit from both Word2vec and BOW information. These hybrid

variants are compared against two baselines given by the classical BOW and Word2vec methods applied individually.

Several tests in the context of sentiment classification are performed on five publicly available Amazon datasets, containing the users' opinions on products coming from different categories [10,11]. The results of our experiments highlighted the usefulness of the novel hybrid representations across the different domains. In particular, the features obtained by concatenating the BOW model with the averaged form of Word2vec consistently outperformed the corresponding baselines.

The remainder of the paper is organized as follows. Section 2 describes the original BOW and Word2vec approaches and the proposed hybrid variants. Section 3 illustrates the classification results achieved on the benchmark datasets. Finally, Sect. 4 contains the conclusions and the future research developments.

2 Representation of Sentences

To use machine learning algorithms the corpus of sentences must be transformed into a rectangular matrix of numeric values. Unlike the BOW model which produces a set of vectors which can be fed directly to a classifier, as described below, word vectors must be manipulated to be converted into unique sentence-level vectors of the same size. When word vectors follow the principle of compositionality, as in the case of Word2vec, two widely used strategies can be adopted. The first simply computes the average of the word vectors in the sentence [17]. The second resorts to a weighted average by considering the TF-IDF (term frequency times inverse document frequency) value of each term [6], which expresses the relative importance of a word inside a sentence [8].

2.1 Bag of Words

Given a corpus defined as a collection $D = (s_1, s_2, \ldots, s_m)$ of m sentences, the BOW model maps each s_i into a numeric vector of dimension V, where V is the size of the vocabulary extracted from D in the form of a set $(v_1, v_2, \ldots, v_V)$ of unique different terms. This vector can be built according to a presence-based approach or, alternatively, to a frequency-based one [15]. In the first case, a given sentence s_i is converted into a boolean vector $\boldsymbol{bow01}_i$ whose generic element j takes the value 1 if and only if v_j appears in the sentence. In the second case, s_i is converted into a vector $\boldsymbol{bowTF}_i$ containing at position j the TF-IDF value of word v_j, defined as the product between the term frequency, i.e. the number of occurrences of v_j in s_i, and the inverse term frequency, given by the logarithm of the ratio of the number of sentences divided by the number of sentences in the corpus containing v_j. Regardless the approach used, D is therefore transformed into a rectangular matrix consisting of V columns and m rows, each one composed by the bag of words of the corresponding sentence.

2.2 Word2vec

The most straightforward way to build a sentence-level vector using Word2vec is to average the word vectors of the terms therein included.

Formally, let $\boldsymbol{w}_{ij}$ denote the d-dimensional pre-trained word vector corresponding to the j^{th} term in sentence s_i. In the averaged Word2vec model, s_i is mapped into a d-dimensional vector $\boldsymbol{avg}_i$ whose k^{th} element is defined as

$$\boldsymbol{avg}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{w}_{ijk} \quad k = 1, 2, \ldots, d, \tag{1}$$

where n_i is the number of words in sentence i.

An alternative approach is represented by the weighted average Word2vec, in which each sentence s_i is converted into the vector $\boldsymbol{avgTF}_i$, whose element at position k is given by

$$\boldsymbol{avgTF}_{ik} = \frac{1}{n_i} \sum_{j=1}^{n_i} f_{ij} \boldsymbol{w}_{ijk} \quad k = 1, 2, \ldots, d, \tag{2}$$

where f_{ij} is the TF-IDF value of word j in sentence i. Notice that, for both strategies the vector representing each sentence has the same dimension d of the input word vectors. As a consequence, the original corpus is mapped into a numeric matrix of size $m \times d$ which can be fed to any machine learning classifier.

2.3 Hybrid Features: Combining BOW with Word2vec

The hybrid sentence representations proposed in this paper draw inspiration from previous studies which analyzed, even if with some limitations, the usefulness of using BOW with Word2vec. The complementarity of the information encoded by the two models, in particular, was empirically highlighted in [4] where an ensemble classifier built on BOW and Word2vec achieved the best results in 9 of the 11 domains for the purpose of sentiment classification. The work of [6] further confirmed this evidence by showing that concatenating TF-IDF weighted average features with frequency-based BOW vectors can outperform other approaches based on the latter alone even if not in all cases.

Along this path and in order to enrich the studies conducted so far, in our work we focused on features engineering rather than ensemble methods such in [4] since they present the drawback of high computational costs. In addition, we also designed novel hybrid embedding variants resorting to different types of vectors concatenation which, at the best of our knowledge, haven't been tested yet. Unlike [6], we introduced the use of a presence-based approach as suggested in [15] and the simple average, which showed better performances, compared to the weighted approach, as reported in [4]. Specifically, by combining in all possible ways the BOW and Word2vec schemes described above, we derived the following four sentence representations

$$avg_i \oplus bow01_i \tag{3}$$

$$avg_i \oplus bowTF_i \tag{4}$$

$$avgTF_i \oplus bow01_i \tag{5}$$

$$avgTF_i \oplus bowTF_i \tag{6}$$

where the operator $\oplus$ denotes the concatenation of vectors. Notice that for each hybrid encoding the vector representing sentence i has a final dimension equal to the sum of the sizes of the component vectors. As an example, if $\boldsymbol{avg}_i$ is a d-dimensional vector and $\boldsymbol{bow01}_i$ has the same size V of the vocabulary of terms extracted from the corpus, $\boldsymbol{avg}_i \oplus \boldsymbol{bow01}_i$ is a $(d+V)$-dimensional feature vector.

3 Experimental Design and Results

To evaluate the effectiveness of the hybrid variants we performed several experiments on a publicly available[1] corpus of Amazon reviews [10,11], referred to products from different categories such as Beauty, Video Games, Clothing, Health and Home. Each review is described by different features: among these, we extracted the review text and its rating, originally ranging from 1 to 5 stars. For the purpose of classification we first discarded the reviews with neutral rating (3 stars) and assigned a positive (negative) polarity to the remaining comments rated more (less) than 3 stars. The polarity was then taken as the binary target variable to predict. A summary of the datasets used in our tests in terms of number of reviews, positive and negative comments, average length of the reviews and size of the vocabulary, is provided in Table 1. Notice that these datasets where chosen since their intrinsic diversity allowed to analyze the performances of the novel encoding schemes across different text domains.

Table 1. Amazon reviews datasets.

Dataset	N. reviews	Positive	Negative	Avg. length	Vocabulary
Beauty	176,229	154,250	21,979	89	14,105
Video games	203,463	174,954	28,509	204	29,201
Clothing	248,230	221,578	26,652	60	11,656
Health	313,057	279,764	33,293	93	17,706
Home	506,423	455,049	51,374	96	15,654

Before the experiments common text data preprocessing in the form of lower case conversion, tokenization and stopwords removal was applied. On the contrary, stemming and lemmatization were not used since, based on a preliminary exploration, these tasks turned out to be ineffective for the specific analysis.

[1] http://jmcauley.ucsd.edu/data/amazon/.

Computational tests were performed on seven different sentence representations, described in Table 2, composed by three baselines and four newly introduced hybrid variants. In particular, the *bow*01 encoding, referred to the presence-based BOW model, was obtained by retaining the first 1000 most frequent terms for each class. The *avg* and the *avgTF* mappings, corresponding to the basic Word2vec schemes, were instead generated by using the 300-dimensional word vectors generated by [12] and trained on a subset of the Google News dataset[2]. Specific terms for which the word vector representation is not available were discarded. Finally, the remaining encodings were built by concatenating the above vectors as described in Sect. 2.3. The code repository is open source and is fully available on request.

Table 2. Alternative sentence representations.

Name	Description	N. features	Type
*bow*01	Presence-based bag of words	1000	baseline
avg	Averaged word vectors	300	baseline
avgTF	Weighted average of word vectors	300	baseline
$avg \oplus bow01$	Concatenate avg and bow01	1300	hybrid
$avg \oplus bowTF$	Concatenate avg and bowTF	1300	hybrid
$avgTF \oplus bow01$	Concatenate avgTF and bow01	1300	hybrid
$avgTF \oplus bowTF$	Concatenate avgTF and bowTF	1300	hybrid

To discriminate between positive and negative reviews we resorted to the Logistic Regression classifier implemented in Weka [19]. Specifically, for each dataset we randomly extracted a stratified training sample of 5000 comments, using the remaining reviews for testing. The most promising ridge regularization parameter, searched among the powers of 10 in the interval $[10^{-4}, 1]$, was obtained through a ten-fold cross-validation on the training set. Furthermore, given the high unbalance of the datasets in terms of class distribution, we selected as performance measure the F1-score on the minority (negative) class.

The results of our experiments are shown in Table 3, which indicates the F1-score computed on the different test sets. The first main outcome is the notable performance exhibited by the hybrid encoding variants compared to the classical BOW and Word2vec representations. In particular, the $avg \oplus bow01$ mapping obtained the highest accuracy across all the datasets. This empirical achievement emphasizes the benefits stemming from the joint use of BOW and Word2vec, suggesting that exploiting the information encoded in the two schemes by concatenating the corresponding sentence-level vectors outperforms the baseline approaches. Notice that, due to the large size of the test sets, it is easy to observe that the 95% confidence intervals around the F1-scores do not overlap. This implies the statistical significance of our results.

[2] https://code.google.com/archive/p/word2vec/.

Moreover, among the hybrid embeddings the one relying on the averaged Word2vec and on the presence-based BOW model consistently provided better results compared to the TF-IDF weighted alternatives. This evidence confirms that the weighting scheme based on TF-IDF, originally proposed to account for the importance of the words in a document within a corpus, is not always an effective choice in a sentiment analysis task.

Table 3. F1-scores on the test sets.

Representation	Beauty	Video games	Clothing	Health	Home
$bow01$	*0.49*	*0.53*	0.51	*0.41*	*0.47*
avg	0.47	0.52	*0.52*	0.37	0.45
$avgTF$	0.38	0.40	0.42	0.29	0.35
$avg \oplus bow01$	**0.51**	**0.54**	**0.52**	**0.43**	**0.48**
$avg \oplus bowTF$	**0.51**	**0.54**	0.51	**0.43**	0.47
$avgTF \oplus bow01$	0.50	0.53	0.50	0.42	0.45
$avgTF \oplus bowTF$	0.50	0.53	0.50	0.42	0.45

Finally, as a further result we observed that the simple $bow01$ encoding generated better predictions compared to both Word2vec baseline schemes on all datasets except for Clothing. Looking at the properties of this dataset we noticed that it collects the shortest reviews, on average, and gives rise to the smallest vocabulary. These two dimensions deserve further investigation since they might play a prominent role on the performance of the Word2vec encodings in different text classification domains.

4 Conclusions and Future Developments

Vectorial embedding of sentences can encode different information about the texts they represent. The most common sentence encoding schemes in the context of machine learning for sentiment analysis are the bag of words (BOW) and the Word2vec models. In the present study we evaluate the usefulness of combining BOW and Word2vec by designing novel hybrid sentence representations which exploit the information provided by both strategies. Experiments on benchmark datasets, which collect the reviews of Amazon's products, confirmed the effectiveness of the hybrid mappings which showed notable performances in terms of prediction accuracy compared to the BOW and Word2vec approaches applied individually. Our empirical finding supports the results obtained in previous studies which conjectured on how the information provided by the well-established BOW scheme can be completed and enriched by the one contained in the more recently proposed Word2vec models. Given the promising results achieved, the present work could be developed in several directions. First, it

would be worthwhile to explore the accuracy of the proposed hybrid variants on a wider collection of textual datasets. Moreover, experiments could be extended by considering alternative classifiers, to evaluate the robustness of the conclusions to the change of the algorithm used for prediction. Finally, other forms of text embedding derived by the combination of numeric sentence encodings could be designed and investigated.

References

1. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
2. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. J. Comput. Sci. **2**(1), 1–8 (2011). https://doi.org/10.1016/J.JOCS.2010.12.007
3. Collobert, R., Weston, J.: A unified architecture for natural language processing. In: ICML 2008, pp. 160–167. ACM Press (2008). https://doi.org/10.1145/1390156.1390177
4. Enríquez, F., Troyano, J.A., López-Solaz, T.: An approach to the use of word embeddings in an opinion classification task. Expert Syst. Appl. **66**, 1–6 (2016). https://doi.org/10.1016/j.eswa.2016.09.005
5. Jansen, B.J., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: tweets as electronic word of mouth. J. Am. Soc. Inf. Sci. Technol. **60**(11), 2169–2188 (2009). https://doi.org/10.1002/asi.21149
6. Lilleberg, J., Zhu, Y., Zhang, Y.: Support vector machines and Word2vec for text classification with semantic features. In: 2015 ICCI*CC, pp. 136–140. IEEE, July 2015. https://doi.org/10.1109/ICCI-CC.2015.7259377
7. Liu, B.: Sentiment Analysis. Cambridge University Press, Cambridge (2015). https://doi.org/10.1017/CBO9781139084789
8. Manning, C.D., Raghavan, P., Schutze, H.: Scoring, term weighting, and the vector space model. In: Introduction to Information Retrieval, pp. 100–123. Cambridge University Press (2008). https://doi.org/10.1017/cbo9780511809071.007
9. Mäntylä, M.V., Graziotin, D., Kuutila, M.: The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. Comput. Sci. Rev. **27**, 16–32 (2018). https://doi.org/10.1016/J.COSREV.2017.10.002
10. McAuley, J., Pandey, R., Leskovec, J.: Inferring networks of substitutable and complementary products. In: ACM SIGKDD 2015, pp. 785–794. ACM Press, New York (2015). https://doi.org/10.1145/2783258.2783381
11. McAuley, J., Targett, C., Shi, Q., van den Hengel, A.: Image-based recommendations on styles and substitutes. In: SIGIR 2015, pp. 43–52. ACM Press, New York (2015). https://doi.org/10.1145/2766462.2767755
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013). http://arxiv.org/abs/1301.3781
14. Mikolov, T., Yih, W.t., Zweig, G.: Linguistic regularities in continuous space word representations. In: NAACL HLT 2013, pp. 746–751 (2013). http://www.aclweb.org/anthology/N13-1090

15. Pang, B., Lee, L.: Opinion Mining and Sentiment Analysis, vol. 2. Now Publishers, Inc., Delft (2008). https://doi.org/10.1561/1500000011
16. Piryani, R., Madhavi, D., Singh, V.: Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. Inf. Process. Manag. **53**(1), 122–150 (2017). https://doi.org/10.1016/J.IPM.2016.07.001
17. Socher, R., et al.: Recursive deep models for semantic compositionality over a sentiment treebank. In: EMNLP 2013, pp. 1631–1642. ACL (2013)
18. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Election forecasts with Twitter. Soc. Sci. Comput. Rev. **29**(4), 402–418 (2010). https://doi.org/10.1177/0894439310386557
19. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc., Amsterdam (2016)

Unintended Bias in Misogyny Detection

Debora Nozza*
University of Milano - Bicocca
Milan, Italy
debora.nozza@unimib.it

Claudia Volpetti*
Politecnico di Milano
Milan, Italy
claudia.volpetti@polimi.it

Elisabetta Fersini
University of Milano - Bicocca
Milan, Italy
elisabetta.fersini@unimib.it

ABSTRACT

During the last years, the phenomenon of hate against women increased exponentially especially in online environments such as microblogs. Although this alarming phenomenon has triggered many studies both from computational linguistic and machine learning points of view, less effort has been spent to analyze if those misogyny detection models are affected by an unintended bias. This can lead the models to associate unreasonably high misogynous scores to a non-misogynous text only because it contains certain terms, called *identity terms*. This work is the first attempt to address the problem of measuring and mitigating unintended bias in machine learning models trained for the misogyny detection task. We propose a novel synthetic test set that can be used as evaluation framework for measuring the unintended bias and different mitigation strategies specific for this task. Moreover, we provide a misogyny detection model that demonstrate to obtain the best classification performance in the state-of-the-art. Experimental results on recently introduced bias metrics confirm the ability of the bias mitigation treatment to reduce the unintended bias of the proposed misogyny detection model.

CCS CONCEPTS

• **Social and professional topics** → **Hate speech**; • **Computing methodologies** → *Neural networks*.

KEYWORDS

misogyny detection, bias measuring, bias mitigation, deep learning

ACM Reference Format:
Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended Bias in Misogyny Detection. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19), October 14–17, 2019, Thessaloniki, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3350546.3352512

1 INTRODUCTION

In the latest years, there was a growing interest in accelerating progress for women's empowerment and gender equality in our society. However, misogyny as a form of hate against them spread exponentially through the web and at very high-frequency rates, especially in online social media, where anonymity or pseudo-anonymity enables the possibility to afflict a target without being recognized or traced. This alarming phenomenon has triggered many studies related to the problem of abusive language recognition, and in particular for misogyny detection, both from computational linguistics and machine learning points of view. However, when inducing a supervised model to perform abusive language classification, it is important to focus on a particular error induced by the training data, i.e. the bias introduced in the model by a set of *identity terms* that are frequently associated to the misogynous class. For example, the term *women*, if frequently used in misogynous messages, would lead most of the supervised classification models to associate an unreasonably high misogynous score to clearly non-misogynous text, such as "You are a woman".

This behavior of recognition models is known as *unintended bias*. In particular, *"a model contains an unintended bias if it performs better for comments containing some particular identity terms than for comments containing others"* [10]. Tackling this error means being able to use those models in the real world.

In this paper, we provide a model for misogyny detection which demonstrates to obtain the best classification performance in the state-of-the-art and we address the fairness of this model by measuring and mitigating its unintended bias. In particular, to address this challenge we first propose a novel synthetic template that can be used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems. Additionally, we investigate different bias mitigation strategies, obtaining a *debiased* model that is less sensitive to identity terms as long as able to perform at the state of art of the best misogyny detection model in the literature on benchmark datasets.

Following, Section 2 provides an overview of the research works for the misogyny detection task and for the bias analysis. Then, Section 3 describes the generation process of the synthetic template test set and the investigated bias mitigation strategies. The evaluation results of the models on several recently proposed bias metrics are reported in Section 4. Finally, in Section 5, conclusions and future work are outlined.

2 RELATED WORK

The state-of-the-art of automatic misogyny identification in online environments is still in its infancy. A preliminary exploratory analysis of misogynous language in online social media has been presented in [17], where the authors collected and manually labeled a set of tweets as positive, negative and neutral, providing some basic statistics about the usage of some candidate misogynistic keywords. A first contribution to the problem of automatic misogyny identification has been presented in [2], where the role of different

*Both authors contributed equally to this research.

WI '19, October 14–17, 2019, Thessaloniki, Greece

ACM ISBN 978-1-4503-6934-3/19/10…$15.00
https://doi.org/10.1145/3350546.3352512

linguistic features and machine learning models have been investigated. More recently, thanks to the Automatic Misogyny Identification (AMI) challenges organized at IberEval [12], Evalita [13], and SemEval [4], many different approaches [3, 14, 15, 19, 20, 22] have been proposed for addressing this problem. In this context, research works commonly focus on textual feature representation studying different linguistic characteristics, ranging from pragmatical, syntactical and lexical features to higher level features derived through embedding techniques, or on the machine learning model, employing traditional or Deep Learning supervised models.

While these works focused on obtaining the most promising performance for the misogyny detection task, they do not explicitly address any study on unintended bias in their misogyny detection models. Addressing biases in text classifiers is crucial, not only because of the potentially discriminatory impact of machine learning models in real-world applications but also because bias correction can improve their robustness when used on different datasets. The research work on bias analysis can be mainly distinguished in two affiliated goals: *measuring* and *mitigating* bias.

Significant recent studies have been published on providing new metrics to quantify the presence of unintended bias in text classification models. Park et al. [21] introduce a measure of the false positive and false negative *Error Rate Equality Differences*, as a relaxation of the equalized odds fairness constraint presented in [16]. These metrics are conceived for binary labels and consequently they strictly depend on the threshold values used to separate the model output scores in two classes. In order to overcome this limitation, Dixon et al. [10] introduce a threshold agnostic metric for unintended bias called Pinned AUC, which has been proven to be inadequate in a follow-up work by the same authors [6]. Consequently, Borkan et al. [7] propose a new set of metrics differing from these early approaches because they are *(i)* threshold agnostic, *(ii)* robust to class imbalances in the dataset, and *(iii)* provide more nuanced insight into the types of bias present in the model. All the metrics cited above will be briefly introduced in Section 4.1.
On the other hand, also bias mitigation in text classification models has been significantly explored recently in the literature. Significant works [5, 10, 11, 16, 21] provide debiasing techniques ranging from debiasing word embedding to data augmentation and fine-tuning data with a larger corpus.

Our work is the first attempt to measure and mitigate unintended bias in misogyny detection models. We provide a state-of-the-art model and we test it against the most recently proposed bias metrics. Finally, we build a debiased version of our model by following the work in [10].
Moreover, since unintended bias cannot be measured on the original test set, debiasing techniques need synthetic unbiased test sets to be generated on purpose for detecting a specific bias. Previous works, such as Kiritchenko and Mohammad [18] and Park et al. [21] generated synthetic datasets for detecting gender bias. Following the identity term template method proposed in Dixon et al. [10], we also provide a novel synthetic template that can be used as the evaluation benchmark dataset for measuring unintended bias in misogyny detection task in future works and that is available online[1].

Class	Train	Test
misogynous	1,785 (45%)	460 (46%)
non-misogynous	2,215 (55%)	540 (54%)

Table 1: Dataset class distribution.

3 METHODOLOGY

3.1 Dataset

In our work, we consider the state-of-the-art corpus for misogyny detection in the English language proposed for the Automatic Misogyny Identification shared task at the Evalita 2018 evaluation campaign [12]. The corpus comprises 4,000 and 1,000 tweets for the training and test set respectively, which has been labeled by human annotators through the Figure Eight[2] crowdsourcing platform. The summary of the class distribution in the corpus is given in Table 1.

3.2 Identity Term Bias

In this paper, the problem of *unintended bias* is addressed by referring to the definition given by Dixon et al. [10].

Definition 3.1. A model contains unintended bias if it performs better for comments containing some particular identity terms than for comments containing others.

This means that despite a misogyny detection model should be biased on misogynistic contents, it should not classify as misogynous tweets that explicitly refer to women or which contains women-related terms only because these are terms that usually appears in misogynistic contents. Indeed, for this study, the identity terms will be terms that can be used to refer to women, which may be unreasonably classified as misogynous with high scores.

Identity Term List. In order to define the list of identity terms, we take into consideration all the synonyms for "woman" by using a thesaurus[3]. The obtained list of synonymous has been then extended by including their plural form. Since some terms (e.g. *gentlewoman*) barely appear in the corpus, we decided to remove the ones with a frequency lower than 3. This choice has been made in order to study the behavior of the misogyny detection model with respect to terms that are actually seen during the training phase. The classification of instances containing identity terms that do not appear in the training set may be influenced by other factors, such as the employed sentence encoding model, exposing the unintended bias analysis to a more complex multifaceted problem which is left to future research.

Identity Term Templates. Since unintended bias of identity terms cannot be measured on the original test set due to class imbalance and highly different identity term contexts, *synthetic test sets* are needed to be generated on purpose.

[1] https://github.com/MIND-Lab/unintended-bias-misogyny-detection
[2] www.figure-eight.com/
[3] www.thesaurus.com

Table 2: Template examples.

Template Examples	Label
<identity_term>should be protected	Non-Misogynous
<identity_term>should be killed	Misogynous
appreciate <identity_term>	Non-Misogynous
hit <identity_term>	Misogynous
amazing <identity_term>	Non-Misogynous
filthy <identity_term>	Misogynous

Following previous work [10], we manually created a balanced synthetic dataset of misogynous and non-misogynous contents. We defined several templates that are filled with the previously identified identity terms and with verbs and adjectives which are divided into negative (e.g. hate, inferior) or positive (e.g. love, awesome) forms to convey hate speech or not. Table 2 reports examples of templates[4]. The generated synthetic dataset comprises 1,464 instances, of which 50% misogynous and 50% non-misogynous, where each identity term appears in the same contexts.

3.3 Misogyny Classification Model

With the purpose of studying the unintended bias problem in a misogyny detection model, we first build a machine learning model on the state-of-the-art misogyny corpus proposed in [12]. Then, we analyze it by measuring by using a synthetic dataset specifically designed for this task. Both datasets are introduced in the previous paragraphs. In this section, we provide details on how we designed and trained the model.

The proposed model, which we will refer to as *reference model*, is outperforming the state-of-the-art classification approaches on the misogyny corpus. We first encoded the English sentences using a novel Deep Learning Representation model, the *Universal Sentence Encoder* introduced in Cer et al. [8] built using a transformer architecture [23] and available online[5]. Once constructed the sentence embeddings, we used them as input to a single-layer neural network architecture and trained our USE_T model. To tackle the model variance, we performed 10 training runs of the same model and then we averaged the results. The model USE_T reached a 72% of mean accuracy on the test set, outperforming of two points the 70% accuracy achieved by *hateminers* team [22] ranked first to the shared task on Automatic Misogyny Identification at the Evalita 2018 evaluation campaign [12]. We implemented the model architecture using the Keras framework [9] with TensorFlow backend [1].

Since we are aware of the fact that sentence embeddings can contain biases themselves [8], we envision as future work an extended version of this study aiming to determine to what extent sentence embeddings encoded biases can affect performances in misogyny detection models.

3.4 Bias Mitigation Strategy

After building our reference model as described in the previous paragraph, we created four debiased versions of our USE_T model in order to mitigate its bias. This section provides further details on the bias mitigation methodologies we used.

We adopted different bias mitigation strategies motivated by the successful work by Dixon et al. [10]. The first one consists of mitigating the class imbalance of the identity terms which have the most imbalanced class distributions. After the class distribution of each identity term is computed, additional data is sampled from an external corpus and subsequently combined to the original training set in order to set the class proportions in line with the prior distribution for the overall dataset. Then, the reference model is trained on this debiased set, originating the *Debiased* model. Moreover, we also build the *Debiased_length* model, which is trained on a debiased set where the class balance is obtained also considering tweet length ranges. This permits to establish the model sensibility to the tweet length when dealing with unintended bias.

In order to confirm the benefits of the described bias mitigation procedure instead of a simple data augmentation process, we investigate the addition of randomly sampled data from the external corpus. The size of the additional random set of tweets is the same of the one computed with the aforementioned mitigation procedure. Analogously, we obtained two bias mitigated models called *Random* and *Random_length* model.

With the aim of maintaining the same language distribution of the training set for the additional data, we employed a state-of-the-art corpus for Hate Speech detection on Twitter [24] as external corpus. Tweets in the corpus have been manually annotated as sexist, racist or neither of them with almost perfect agreement. To mitigate the impact of the random sampling, both the procedures are repeated over 10 runs, originating 10 different training sets for each model.

In the following, in order to measure and evaluate our USE_T model bias, we compare it against its *Debiased*, *Debiased_length*, *Random* and *Random_length* debiased versions.

4 EXPERIMENTS

This section briefly describes the investigated metrics and subsequently reports their evaluation on the test set and on the generated synthetic dataset.

4.1 Metrics

We adopted the AUC (area under the curve) measure to evaluate the classification performance of the misogyny detection model on the test set and on the synthetic dataset. Concerning the unintended bias analysis, we computed the metrics introduced in recent state-of-the-art works [7, 10] to measure the extent of unintended bias in the model. The *Error Rate Equality Differences* measures the variation of the false positive and false negative rates between identity terms. The hypothesis motivating these metrics is that a model without unintended bias will have similar error rates across all identity terms. Since Error Rate Equality Differences measures the classification outcomes, and not the real-valued score as AUC, we applied a 0.5 threshold to discriminate between the two classes.

We decided to not investigate the Pinned AUC metric as it has been proved to suffer from several limitations [6] and that its ability to reveal unintended bias is highly impacted by a sampling procedure [10]. As suggested in [10], we investigated three separate

[4]The complete set of identity terms, verbs and adjectives is available at https://github.com/MIND-Lab/unintended-bias-misogyny-detection.
[5]https://tfhub.dev/google/universal-sentence-encoder-large/3

Model	Test	Templates
USE_T	0.7170	0.6339
Debiased	0.7045	0.6423
Random	0.7127	0.6396
Debiased_length	0.7003	0.6437
Random_length	0.7140	0.6376

Table 3: Mean AUC on the test and synthetic templates sets.

AUC-based metrics, recently defined in [7], which provide a more detailed view than Pinned AUC, and thus providing a more general framework for measuring unintended bias.

These metrics are calculated using the score distributions of both the whole background test data and the test set subgroup containing the identity term itself. *Subgroup AUC* (subAUC) metric provides a measure of the separability within the example from the subgroup. *Background Positive Subgroup Negative AUC* (BPSN) metric calculates AUC on the positive examples from the background and the negative examples from the subgroup. If this value is high, then it is likely that fewer negative examples from the subgroup are classified as false positives at many thresholds. *Background Negative Subgroup Positive AUC* (BNSP) metric calculates AUC on the negative examples from the background and the positive examples from the subgroup. If this value is high, then it is likely that fewer positive examples from the subgroup are classified as false negatives at many thresholds. Unfortunately, each metric provides a bias measure on a specific term exclusively. Hence, in order to combine the three per-term AUC-based metrics into one overall bias measure, we calculated their generalized mean and finally their weighted average with the overall model AUC[6], i.e. the *Weighted Bias Score*.

Additionally, two threshold agnostic metrics are studied. *Positive Average Equality Gap* (posAEG) and *Negative Average Equality Gap* (negAEG), as defined in Borkan [7], measure the separability of positive examples from the subgroup with positive examples from the background data and vice-versa. They range from -0.5 to 0.5 and their optimal value is 0. When close to the optimal value, there is no score shift from the subgroup positive examples and the background positive data since the distributions have an identical mean. The combined use of AUC-based metrics and AEGs, provide a detailed view of the types of bias present in the considered model.

4.2 AUC

The performance, in terms of AUC, on the test and synthetic templates sets are reported in Table 3. As a general remark, it is possible to notice that all the employed debiasing techniques have been effective on improving the mean AUC on the Identity Term Templates, while maintaining comparable performance on the test set with respect to the reference model USE_T. Comparing the results obtained with the debiasing and random treatments enables us to demonstrate that the improvements achieved by mitigating the bias are not solely due to the addition of data. The consideration of the tweet length in the bias mitigation phase has been proven to be beneficial for reducing the unintended bias.

[6]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity%2Dclassification/overview/evaluation

Metric	False Positive Equality Difference	False Negative Equality Difference
USE_T	17.49	20.64
Debiased	9.61	18.65
Random	11.44	26.28
Debiased_length	**8.80**	**12.42**
Random_length	12.18	26.90

Table 4: Average of the Error Rate Equality Differences for each model.

4.3 Error Rates

A further investigation on the analysis of unintended bias has been carried out by comparing the false positive and false negative error rates for each identity term of each model considered. It is important to mention that, with the aim of evaluating the bias, it is not important to observe the punctual values of these metrics but rather than they have similar values across all identity terms. This means that the presence of a specific identity term in a tweet is not causing an increase (or decrease) in the error rates and consequently it is not subjected to unintended bias.

Figures 1 and 2 report the false positive and false negative error rates, for each identity term, of the reference model (USE_T) and the models trained after the bias mitigation strategy considering the tweet length. Each point in the chart corresponds to the error rate of each model configuration, indeed USE_T is represented with 10 points and the bias mitigated models by 100.

By looking at false positive rates (Figure 1), it is possible to draw two different conclusions: the bias mitigation strategies, and in particular the non-random one, have (i) significantly decreased the false positive rates for each identity term and (ii) reduced the unintended bias by providing more similar values across terms.

In Figure 2, the false negative rates also demonstrate that the bias mitigation strategies are able to limit the problem of unintended bias by mitigating the differences across terms. Even if it is not essential for the bias mitigation extent, an additional consideration can be made about the absolute values of this measure, which show a different behavior from the false positive rates. In this case, the debiased models obtained higher false negative rates with a high variance among the configurations. This can be probably due to the fact that the bias mitigation strategies are specifically aimed to solve the false positive issues introducing only negative examples. Consequently, as a counter-effect, the model becomes less accurate on classifying negative examples.

4.4 Equality Difference Summary

In order to provide a more immediate comparison between the models, Table 4 reports the results in terms of Error Rate Equality Differences, distinguishing false positive and false negative. These results confirm the considerations made based on Figures 1 and 2, i.e. the bias mitigation strategies are reducing the unintended bias with respect to the reference model USE_T.

In particular, the improvements of the debiased model are even more evident when comparing to the model trained after the random debiasing treatment, demonstrating that the results are not

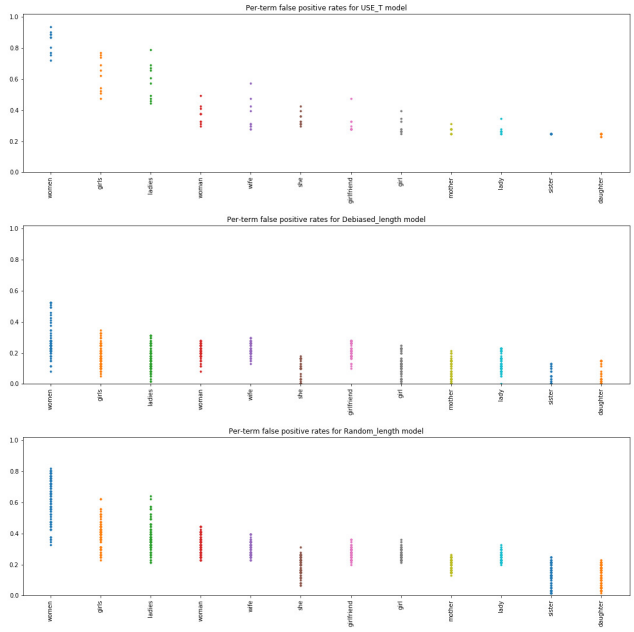

Figure 1: False positive rates for each identity term of the reference and debiased models.

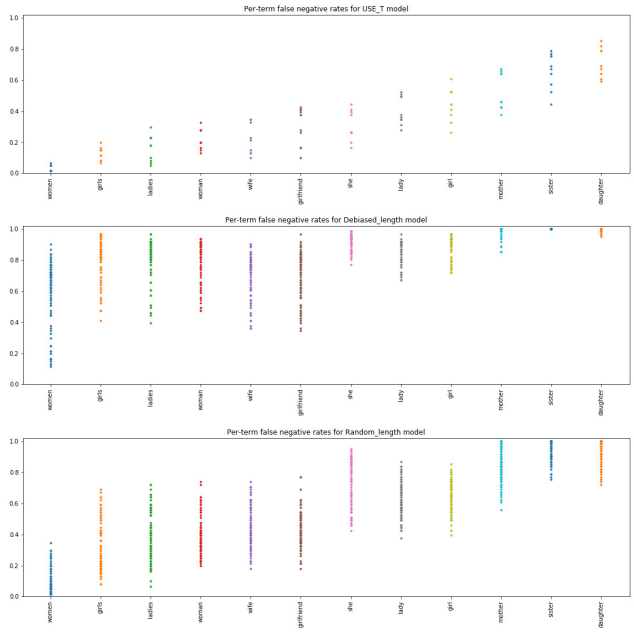

Figure 2: False negative rates for each identity term of the reference and debiased models.

due only to the addition of data. Moreover, it is possible to observe that the consideration of the tweet length in the bias mitigation strategy has lead to better results.

Model	Weighted Bias Score (power mean)	Weighted Bias Score (aritmetic mean)
USE_T	0.594	0.641
Debiased	0.593	**0.654**
Random	0.591	0.646
Debiased_length	**0.595**	0.652
Random_length	0.586	0.644

Table 5: Weighted Bias Scores for each model.

4.5 AUC-based metrics and AEGs

In Figure 3, we report the heatmaps for the full set of AUC-based metrics (subAUC, BPSN, BNSP) and the AEGs (negAEG, posAEG) metrics. All metrics are calculated for each identity term and heatmaps compare the USE_T reference model with the *Debiased_length* model, which demonstrated to be the most effective in reducing unintended biases according to previous analysis.

For the sake of a fair comparison, the heatmaps report the best results for each model across the sampling runs. By examining the results, we can observe that the debiased model shows a stable improvement of the subAUC measure across all terms, confirming a higher separability of positive and negative examples within each subgroup, if compared with the USE_T model subgroups separability. According to the types of biased taxonomy defined in Borkan [7], we can say that our reference model USE_T is likely to suffer from the so-called *wide subgroup score range with overlap* and *low group separability* types of bias. This can be explained by the evidence that (i) it underperformed on most of the subgroups resulting in a lower separability within subgroups compared to the background distribution and (ii) the subgroup scores distributions are so wide that they overlap with each other and with the opposite class background distributions. After the debiasing process has been applied to the model, both types of bias results mitigated, motivated by the fact that the per-subgroups AUCs are finally comparable to the mean AUC of the debiased model (see Table 3). Results in Figure 3 also show an increase in the BPSN measure on nine out of twelve sub-groups, resulting in a reduction of False Positives for the relative identity terms. A similar improvement is reported for the BNSP measure, demonstrating a reduction of the False Negatives for those subgroups that report a higher value for the metrics. Results in terms of AEGs report slight shifts of most of the subgroup distributions caused by the attempt of debiasing, but they are never reduced to their optimal value 0.

Table 5 reports the Weighted Bias Score[7], a summary metric able to combine the overall AUC with the three AUC-based metrics (subAUC, BPSN, BNSP). Debiased models outperform both the random models and the USE_T reference model, demonstrating the ability to reduce the unintended biases without losing in overall performances. A power mean (with p=-5 as suggested by authors metric) and an arithmetic mean are applied and both variants of the Weighted Bias Score results in a higher value for the debiased models with respect to the other models.

[7]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity%2Dclassification/overview/evaluation

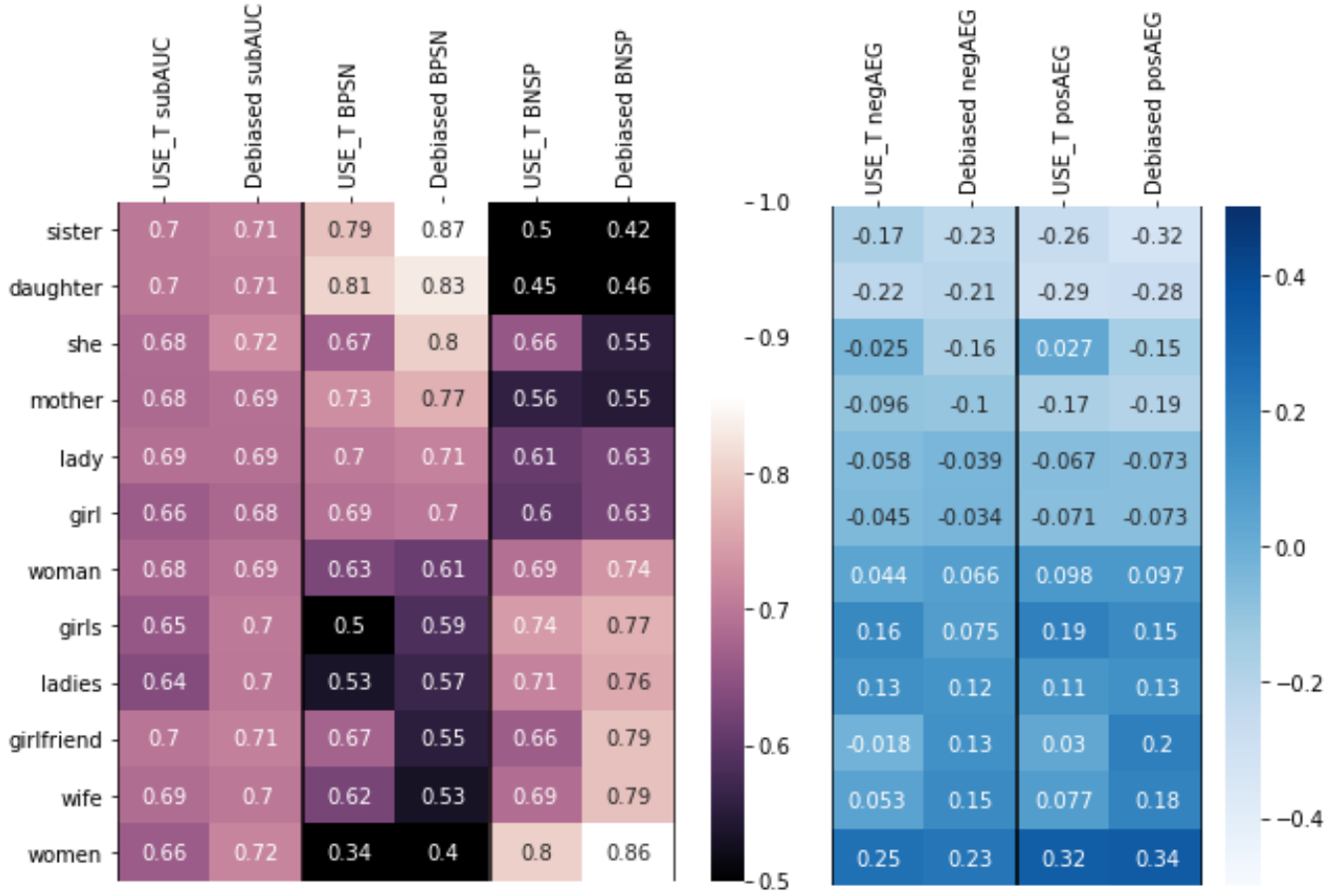

Figure 3: Comparison between USE_T and Debiased model on the synthetic dataset.

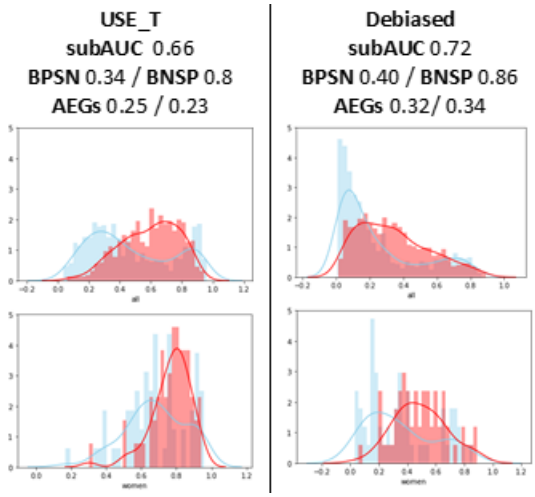

Figure 4: Bias reduction on "women".

Finally, Figure 4 provides an example of the debiasing method impact in reducing unintended biases for one of the most frequent identity terms in our dataset: "women". Plots reported in Figure 4 aim at graphically displaying that the subgroup separability of positive and negative examples for the debiased model is higher then the case of the reference USE_T model. This is demonstrated indeed by the increase in the subAUC value up to 0.72. This reflects on smaller numbers of False Positives and False Negatives misclassified examples. BPSN and BNSP improvements demonstrate the decrease of respectively the overlapping of negative subgroup samples with the positive background and vice-versa. Both AEGs are positives, corresponding to right-shifts of both the score distributions of the subgroup.

5 CONCLUSIONS AND FUTURE WORK

This paper presents the first attempt to address the problem of measuring and mitigating unintended bias in machine learning models trained for the misogyny detection task. We proposed a state-of-the-art model for misogyny detection, based on a transformer architecture, and we studied its unintended bias with some of the most recent metrics in literature.

We investigated different bias mitigation strategies, obtaining a debiased version of the proposed model that is less sensitive to identity terms as long as able to perform at the state of art of the

best misogyny detection model in the literature on benchmark datasets. The bias mitigation strategies have significantly decreased the false positive and false negative rates for each identity term and consequently reduced the unintended bias by providing more similar values across terms. The debiased model showed a stable improvement in separability of positive and negative examples within each subgroup, if compared with the reference model subgroups. Additionally, we first propose a novel synthetic template set that can be used in the future as a benchmark test set for measuring the unintended bias in misogyny detection problems.

As future work, we envision an extended version of this study aiming to determine to what extent sentence embeddings encoded biases can affect performances in misogyny detection models. The idea is to analyze and compare the impact on performances and biases of machine learning models based on pre-trained embeddings against a baseline where embeddings are trained during the learning phase.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015). https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic Identification and Classification of Misogynistic Language on Twitter. In *Proceedings of the 23rd International Conference on Applications of Natural Language to Information Systems (NLDB 2018) (Lecture Notes in Computer Science)*, Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane (Eds.), Vol. 10859. Springer, 57–64.

[3] Angelo Basile and Chiara Rubagotti. 2018. CrotoneMilano for AMI at Evalita2018. A Performant, Cross-lingual Misogyny Detection System.. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

[4] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval 2019)*. Association for Computational Linguistics, 54–63.

[5] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *arXiv preprint arXiv:1707.00075* (jun 2017). arXiv:1707.00075

[6] Daniel Borkan, Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Limitations of Pinned AUC for Measuring Unintended Bias. *arXiv preprint arXiv:1903.02088* (2019).

[7] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In *Companion of The 2019 World Wide Web Conference (WWW 2019)*. ACM.

[8] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP 2018)*. Association for Computational Linguistics, 169–174.

[9] François Chollet et al. 2015. Keras. https://keras.io. (2015).

[10] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 67–73.

[11] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 11–21.

[12] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (AMI). In *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR-WS.org.

[13] Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*. CEUR-WS.org.

[14] Simona Frenda, Bilal Ghanem, and Manuel Montes-y Gómez. 2018. Exploration of Misogyny in Spanish and English tweets. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Vol. 2150. CEUR-WS.org, 260–267.

[15] Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Arantza Casillas, Arantza Díaz de Ilarraza, Nerea Ezeiza, Maite Oronoz, Alicia Pérez, and Olatz Perez-de-Viñaspre. 2018. Automatic Misogyny Identification Using Neural Networks. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*. CEUR-WS.org.

[16] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*. 3315–3323.

[17] Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The Problem of identifying Misogynist Language on Twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*. ACM Press, 333–335.

[18] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM@NAACL-HLT 2018)*. Association for Computational Linguistics, 43–53.

[19] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. 14-ExLab@ UniTo for AMI at IberEval2018: Exploiting lexical knowledge for detecting misogyny in English and Spanish tweets. In *Proceedings of the 3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018)*, Vol. 2150. CEUR-WS.org, 234–241.

[20] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, and Viviana Patti. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Proceedings of 6th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*. CEUR.org.

[21] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. Association for Computational Linguistics, 2799–2804.

[22] Punyajoy Saha, Binny Mathew, Pawan Goyal, and Animesh Mukherjee. 2018. Hateminers: Detecting Hate speech against Women. *arXiv preprint arXiv:1812.06700* (2018).

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems (NIPS 2017)*. 6000–6010.

[24] Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the Student Research Workshop, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (SRW@HLT-NAACL 2016)*. Association for Computational Linguistics, 88–93.

Temporal Word Embeddings for Narrative Understanding

Claudia Volpetti
Politecnico di Milano
Via Lambruschini 4b
20156 Milan (Italy)
claudia.volpetti@polimi.it

Vani K
IDSIA
Galleria 2 - Via Cantonale
Manno - Lugano (Switzerland)
vanik@idsia.ch

Alessandro Antonucci
IDSIA
Galleria 2 - Via Cantonale
Manno - Lugano (Switzerland)
alessandro@idsia.ch

ABSTRACT

We propose temporal word embeddings as a suitable tool to study the evolution of characters and their sentiments across the plot of a narrative text. The dynamic evolution of instances within a narrative text is a challenging task, where complex behavioral evolutions and other characteristics specific to the narrative text need to be inferred and interpreted. While starting from an existing approach to the learning of these models, we propose an alternative initialization procedure which seems to be especially suited for the case of narrative text. As a validation benchmark, we use the Harry Potter series of books as a challenging case study for such character trait evolutions. A benchmark data set based on temporal word analogies related to the characters in the plot of the series is considered. The results are promising, and the empirical validation seems to support the working ideas behind this proposal.

CCS Concepts

• **Artificial intelligence → Natural language processing**
• **Machine learning → Neural networks.**

Keywords

Natural Language Processing; Word Embeddings; Temporal Word Embeddings; Narrative Understanding; Character-Centric Narrative Understanding; Temporal Word Analogies.

1. INTRODUCTION

Narrative Understanding (NU) tasks are natural language understanding techniques specifically designed to process narrative texts and automatically extract from them higher-level information. NU examples are associated to the concepts of narrative storytelling, event chain analysis, narrative generations and inferencing to social media narrative analysis. Efforts in NU are focused on learning the sequence of events by which a story is defined; in this tradition we might situate seminal work on learning procedural scripts [1,2], narrative chains [3], and plot structure [4].

If those works are *story-centric*, i.e., the focus is on the plot of the story, while some other approaches are *author-centric*, i.e., focused instead on plot coherences, here we analyze much more the characters and their relations. *Character-centric* approaches are focused on character believability, i.e., the extent to which the characters in a story exhibit rich and diverse interactions, emotions, social behavior and motivations [5]. *Character-centric NU* (CNU) tasks are therefore methods focused on understanding and exploring such character believability attributes of the narratives from a social perspective. Topics include identifying characters in narratives, modeling characters as social goal-oriented agents, their interaction with other characters or the environment, their similarity with other entities, their evolution over time, and others.

Acting under the CNU umbrella, we consider the task of automatically identifying *characters roles and their evolution* over time. A character *role* describes what function a character serves in the story. The character *evolution* is the idea in writing that a character can ideally change from the beginning of a work to the last sentence, e.g., from the villain to the hero. To validate these techniques, we use J.K. Rowling's *Harry Potter* books as a benchmark providing a consistent amount of text, with a story spread over multiple books with recurrent characters and varying relations among them.

Tackling of such a task inherently demands an integration of natural language processing and advanced machine learning. In our approach to CNU, we use *temporal word embeddings* (TWEs), as a tool to represent the time-varying semantic distributions of a vocabulary. A *word embedding* E is a map from a vocabulary V of size v to a d-dimensional real space, i.e., $E: V \to \mathbb{R}^d$, provided together with a metric $\delta: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, that evaluates the relative distance between these vectors. Given two words $w_1, w_2 \in V$, the nonnegative real number $\delta(w_1, w_2)$ measures the dissimilarity level between the two words [6]. Word embedding training is achieved within neural networks architecture. In the simplest setup, a v-dimensional input layer goes to a d-dimensional hidden layer through a $v \times d$ input weight matrix $\boldsymbol{W}$ (also called *word embedding matrix*) and the hidden layer goes to a v-dimensional output layer through a $d \times v$ output weight matrix $\boldsymbol{W'}$ (also called *context matrix*). Each word of a text together with its neighboring word(s) can be used as a set of input/output data able to train the word-to-word map $\boldsymbol{W} \cdot \boldsymbol{W'}$, and $\boldsymbol{W}$ alone eventually provides the required embedding. TWE models are recently proposed approaches to the dynamic learning of word embeddings, i.e., vectors that represent the meaning of words, during a specific temporal interval. Formally, a TWE $\{E_t\}_{t \in T}$ is just a parametrized set of word embeddings, where the parameter t belongs to a set T, that can be discrete or continuous, and for each $t \in T$, E_t is a word embeeding defined as in the previous paragraph. An example is in [7], a TWE is expected to associate different vectors to the word *gay* at different times in the history: its vector in 1900 is expected to be more similar to *cheerful* than its vector in 2005.

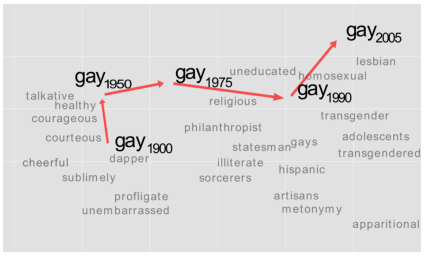

Figure 1 Two-dimensional visualization of semantic change in English using temporal word vectors of the word *gay* [22].

Stemming from this idea, our claim is that if we construct TWEs (also known as *dynamic word embeddings* [8] or *diachronic word embeddings* [7]) for each character in different time periods (e.g., for each character in each book of a book series) they can be used to represent the role and the evolution of a character along a story plot. TWEs make it possible to find distinct words that share a similar meaning in different periods of time by retrieving temporal embeddings that occupy similar regions in the vector spaces that correspond to distinct time periods. Consequently, our hypothesis is that characters having the same role, they are assumed to be closer, according to some metric measure in the embedding space, to similar characters that producing a distribution in the vector space so that, e.g., villains should be clustered in a different area from the area in which story heroes are placed. On the other hand, we also claim that by building a sequence of temporal embeddings of a character over consecutive time intervals, one can track the character evolution (semantic shift) occurred in the character role.

Moreover, in this work, we suggest the use of *Temporal Word Analogies* (TWAs) [9] as a tool to evaluate character evolution, since TWAs are one of the standard approaches to the evaluation of TWEs in general. A TWA holds when two words share a common meaning at two different points in time, e.g., "*Ronald Reagan in 1987 is like Bill Clinton in 1997*". The task is therefore to find the word w^* with the semantic role at time t most similar to that of a word w' at a different time t', i.e.,

$$w' : t' = w^* : t$$

Using TWEs to solve TWAs is based on the implicit idea of an *alignment* the semantic areas in the codomains of the different embeddings associated to a TWE. E.g., an area associated to the *US President* occupied by *Ronald Reagan* vector in 1987 and by *Bill Clinton* vector in 1990. Accordingly, a TWE-based of a TWA is:

$$w^* = \arg min_{w \in V} \, \delta(E_t(w), E_{t'}(w'))$$

Accordingly, we can use TWAs to validate the hypothesis that TWEs of characters can be used for CNU as they provide information on characters roles and evolution. In the considered benchmark, the different books of the series are natural timestamps for the TWE and we consider therefore TWA as the following:

$$Voldemort : Book\ I = ? : Book\ II$$

i.e., who is the character whose role in Book II is more similar to that of Voldemort in Book I. The accuracy in solving such TWAs is therefore a possibly proxy of the effectiveness of adopting TWEs for CNU. To measure such accuracies, we create a data set of TWAs across all the books of the Harry Potter saga, gathered through ten annotators with deep knowledge and understanding of these books. To the best of our knowledge, this is the first work that attempts to learn explicit character roles and their evolution in narratives by TWEs.

The paper is organized as follows. Section 2 summarizes the state of the art in CNU. Section 3 discusses the experimental setup with details on TWAs data sets and our approach to TWE training. Section 4 reports on the experimental results. The paper is concluded in Section 5 with brief insights to future directions.

2. RELATED WORK

Automated story understanding is a long-pursued task for AI [10,11]. This has been approached as a commonsense reasoning task, by which systems make inferences about events that prototypically occur in common experiences [12]. Early works often failed to scale beyond narrow domains of stories due to the difficulty of automatically inducing domain-specific knowledge. The shift to data-driven AI established new opportunities to acquire this knowledge automatically from story corpora. Nowadays natural language processing recognizes that the type of commonsense reasoning used to predict what happens next in a story, for example, is as important for natural language understanding systems as linguistic knowledge itself. Regarding the specific area of CNU, as already mentioned in the introduction, most of the efforts have been in the direction of character identifications and understanding the evolutions on semantic space. This include prediction of event sequences, emotional trajectories [13,14], identification of sentiments and relations [15,16] and generation of character networks and other visualizations [17,18].

3. EXPERIMENTAL FRAMEWORK

We intend to explore the semantic and temporal spaces of characters in the narrative using TWE. In this section, we discuss how we trained TWEs and tested them using a TWAs data set.

3.1 Training Data

When training TWEs, the amount of information we are able to encode is heavily influenced by the type and size of textual data being used for their training and the temporal granularity of the data [9]. For our experiments we considered as a corpus the six books from the Harry Potter's series. Since the training process of a TWEs relies on diachronic text corpora, we need to decompose our corpus into temporal slices [5,10]. Usually, temporal intervals are set accordingly to the granularity of time spans we want to cover with TWEs [9]. Since we are trying to trace major changes in characters behaviors and role, we decided to keep the granularity of time spans low and consequently we set our *time unit* (the granularity of the temporal dimension) to the number of books. As a result of this choice, after the training every word will have six representations, one per each time unit (per each book). Note that we work under the assumption that both the *narrative order* and the *chronological order* of the events and character evolution coincide in the corpus.

Table 1 – Temporal Word Analogies Data Set

Book	Main Antagonist	Second Antagonist	Main Alley	Second Alley	Third Alley
I	*Voldemort*	*Quirrell*	*Ron*	*Hermione*	*Hagrid*
II	*Riddle*	*Basilisk*	*Ron*	*Hermione*	*Hagrid*
III	*Dementors*	*Pettigrew*	*Ron*	*Hermione*	*Lupin*
IV	*Voldemort*	*Crouch*	*Hermione*	*Ron*	*Cedric*
V	*Voldemort*	*Umbridge*	*Ron*	*Hermione*	*Sirius*
VI	*Voldemort*	*Snape*	*Ron*	*Hermione*	*Dumbledore*

3.2 Test Data on Temporal Word Analogies

To test the strength of TWE in CNU we consider a TWAs data set that we built on purpose for this task. This section illustrates how we design and build this dataset. We cope with the Harry Potter's books corpus, and asked ten "experts", i.e., people who carefully and repeatedly read the six books, to answer a survey. They were asked to answer twelve questions about Harry Potter's characters across the first six books. This approach made it possible to trace a series of 150 characters analogies over time. Table 1 reports examples of characters from the first book and their analogues over the following books as gathered from annotators from which TWA ground truth can be obtained. Each column represents a character role, and the names in that column reports the different characters embodying that role at different points in time (i.e., books).

3.3 Character Identification

Characters are a key element of narrative and so character identification is a necessary preprocessing task for the kind of analysis considered in this paper. Named entity recognition tools such as the classic API of the Stanford dependency parse can be used for that. Yet, unlike other kind of texts, in the particular case of NU, the same character might often appear with different *aliases* (e.g., *Harry Potter* as *Harry*, *Ronald Weasley* as *Ron* or *Ronald*). The model we want to develop should clearly work at the character level and the different aliases of a same character need to be regarded as a single word vector. A clustering procedure might be achieved by standard techniques from unsupervised learning techniques (e.g., DBSCAN) with a set of additional heuristic rules.

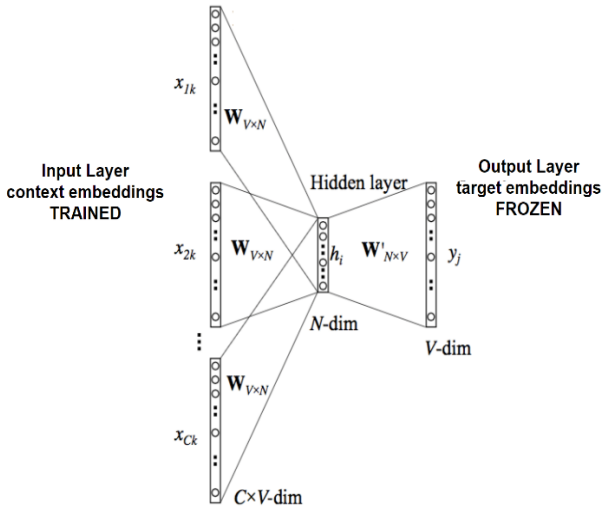

Figure 2 – Training input vectors with frozen output vectors.

3.4 Training Temporal Word Embeddings

Recently, different researchers have been demonstrating the efficiency of tracing temporal changes in lexical semantics using an approach known as distributional models. Such models seem well suited for monitoring the gradual process of meaning change of words over time. Several recent publications demonstrated these models to be efficient and outperform the frequency-based methods in detecting semantic shifts of words over time [7, 22]. In particular we focus on the case of TWEs.

Many training methods for TWEs suffer from *alignment* issues, i.e., once you train separate word embedding at different time periods (on different corpus slices), it does not make sense to directly calculate similarities between vectors of one and the same word in two different time periods. This is related to the inherent stochasticity of most word embedding training algorithms. To solve this, [22] suggested to first align the models and the calculating similarities. Yet, it has been shown that alignment can compromise the information encoded in the embeddings.

The specific method introduced in [19] seems instead to be able to implicitly align different temporal representations using a shared coordinate system instead of enforcing vector similarity in the alignment process. The same model also proved to be easy to implement on the top of continuous bag of words and skip-grams as Word2vec architecture and highly efficient to train.

This method is built on the assumption that a word, e.g., *Clinton* appears during some temporal periods in the contexts of words that are related to his position, e.g., *president*, that conversely doesn't change its meaning. This assumption allows to heuristically consider the context matrix as static, i.e., to freeze the output weight matrix during training, while allowing the word embedding input weight matrices, to change on the basis of co-occurrence frequencies that are specific to a given temporal interval (Figure 2). After training, model returns the context embeddings, that we are going to consider as a TWE.

This is achieved by a two-fold training procedure. First a static word embedding is trained, with random initialization, using the entire vocabulary and ignoring temporal slices. Let us denote as $\boldsymbol{W}$ the corresponding word embedding matrix and as $\boldsymbol{W'}$ the corresponding context matrix. The word embedding matrices of the TWE, say $\{\boldsymbol{W}_t\}_{t\in T}$, is achieved by initializing these matrices with $\boldsymbol{W}$ and keeping $\boldsymbol{W'}$ as a *frozen* context matrix equal for all the time slices. This initialization has been proved to force alignment and make it possible to compare vectors from embeddings associated to different time slices. Note also that the same procedure with $\boldsymbol{W}$ frozen and $\boldsymbol{W'}$ as initialization could be considered.

In this paper we propose a different initialization scheme for such a training architecture. Having $\boldsymbol{W}$ as the same initialization for all the word embeddings associated to different time slices reflects the idea of a common background of *semantically static* words, which are practically not changing their meaning over time. In the particular case of the characters of a narrative text the situation might be different, as basically each character might change their semantic position over time. For this reason, a better initialization strategy might consist in using $\boldsymbol{W}_{t-1}$ as the initialization of $\boldsymbol{W}_t$ and so on, while using $\boldsymbol{W}$ only for the model of the first slice $\boldsymbol{W}_0$. We call this procedure *dynamic initialization*, while the original procedure proposed in [19] is called here *static initialization*. A graphical summary of the architecture together with the two initialization strategie is depicted in Figure 3.

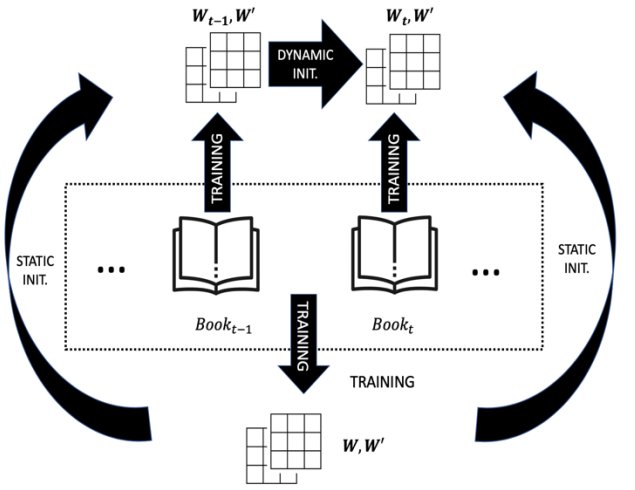

Figure 3 - Temporal context embeddings architecture with both static and dynamic initialization.

Conference'10, Month 1–2, 2010, City, State, Country.

DOI: http://dx.doi.org/10.1145/12345.67890

4. RESULTS AND ANALYSIS

In this section, we discuss the details of our experimental results obtained from the baseline model [19] and the variant with dynamic initialization we proposed in this paper. This is achieved by means of an implementation of the continuous bag of words and the negative sampling extending the Gensim library.[1]

More specifically, we trained both models on the entire Harry Potter corpus (six books) to build the static embeddings and then we trained separately the temporal embeddings according the two different approaches discussed in the previous paragraph. After some hyperparameters tuning procedures, we fixed $d = 200$ for the word embeddings dimensionality, we specified a window size equal to two, and twenty epochs for the static training. Temporal embeddings were trained for five epochs each, when replicating the original approach, and gradually scaled from ten to one when dealing with the second approach.

4.1 Temporal Embeddings Visualization

Semantic trajectories, consisting in the set of vectors corresponding to the same word over different times are the most straightforward product provided by a TWE. Standard techniques such as t-SNE[2] can be used to project the high-dimensional vector to spaces of dimention two or three and visualize the semantic trajectories as in the examples in Figure 1 or in Figure 4.

Here are going to plot the temporal embeddings resulting from the training. We are going to discuss the ones produced by the original approach here in the following, since both approaches provided very similar visualizations.

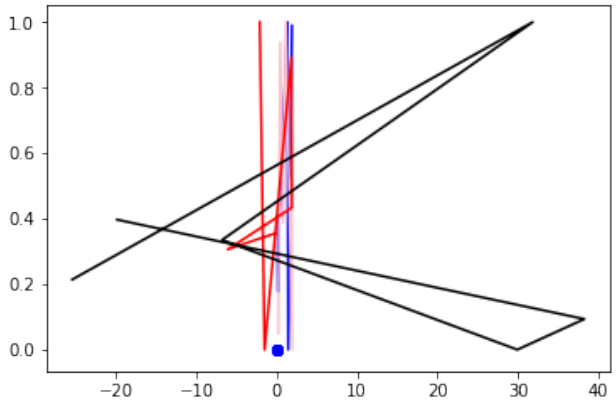

Figure 4 – Semantic trajectories across the six books of five characters: Harry (blue dot), Ron (blue line), Hermione (pink line), Draco (red line) and Voldemort (black line).

In Figure 4, the fixed point is representing Harry and the lines are reporting the behavior through time of the other characters. Hermione and Ron cover the same role in the story plot, and both are main Harry's alleys, figure shows clearly that they both follow the same path, since their lines are close and behave similarly. Voldemort is the main antagonist through books, as it is clearly depicted in figure, its line follows a different path far from the alleys and occupy very different areas in space. Interestingly different is the Draco's vector behavior. Draco character ranges from serving as a secondary antagonist to supporting antagonist, to finally being the central antagonist far ahead in the story plot. This more complex evolution of the character can be confirmed by observing the character evolution line in figure. It is a path which is close to Ron and Hermione behaviors but has a lot of traits of the main antagonist (Voldemort) path.

In summary, Figure 4 seems to demonstrate that TWEs can be used to plot character evolutions and identify character roles by observing their movements and positions in the vector space.

4.2 Temporal Word Analogies Results

From an implementational point of view, solving a TWA is just a matter of retrieving the temporal vector of character in a particular book, and then finding the closer point to that vector among all vectors in a second book as in the equation in Section 1. The resulting vector will be the solution of the analogy. Following this procedure, we can find the characters in a second book most similar to a certain character in another book. Given the TWAs dataset we introduced, we used our models to predict the correct results of 150 temporal character analogies.

Table 2 – Example of Temporal Model Predictions

Antagonist book$_A$	**Antagonist book$_B$**	**Prediction book$_B$**
Voldemort	**Riddle**	**Riddle** ✓
Quirrell	Basilisk	Snape ✗
Ron	**Ron**	**Ron** ✓
Hermione	**Hermione**	**Hermione** ✓
Hagrid	Hagrid	Colin ✗

Accuracy is chosen as metric to count how many correct analogies are predicted. It simply counts the number of correct predictions divided by the total number of analogies. We also provide the accuracy calculated for the *static* and *dynamic* analogies separately (Figure 5). Static analogies involve the same word. E.g., *Voldemort: Book1* = *Voldemort: Book3* is static one, while dynamic analogies involve different words.

In Table 2 we report an example of predictions for both static and dynamic analogies. You can interpret the table as follows: e.g. first row is a dynamic analogy since the characters involved are different and the prediction of our model is in this case correct; firth row reports a static analogy and in this case our model output the wrong prediction.

The results of the experiments are summarized in Figure 5. Both models reached very similar performances in terms of general accuracy. We should highlight the fact that both models don't have any difficulty in predicting static analogies (both reach more than 99% of accuracy) and that our variant performs slightly better when facing dynamic analogies.

TWE Model	**Accuracy**	**Static**	**Dynamic**
Static Initialization	65.07 (97.6/150)	99.63 (53.8/54)	45.62 (43.8/96)
Dynamic Initialization	65.14 (97.7/150)	99.26 (53.6/54)	45.94 (44.1/96)

Figure 5 – Accuracy performances on temporal word analogies data set in case of all, only static and only dynamic analogies.

[1] https://github.com/valedica/twec

[2] https://scikit-learn.org/

Finally, in order to be sure to use a fair metric, we calculated a different type of metric as alternative to the one in Figure 5. The accuracy metric used so far, was designed to consider not only the first prediction, but the five top closer vectors predicted as similar characters by using a weighted sum of the errors. In Figure 6, we also provide an alternative accuracy results including only the top 2 predictions. In this case, again models' performances are comparable, and we also record a slightly better outcome for the original model.

TWE Model	Accuracy (Top 5)	Accuracy (Top 2)
Static Init.	65.07 (97.6/150)	54.8 (82.2/150)
Dynamic Init.	65.14 (97.7/150)	54.3 (81.4/150)

Figure 6 - Accuracy performances on TWAs benchmark.

5. CONCLUSIONS AND OUTLOOKS

We studied temporal word embeddings as a possible tool for effective character-centric narrative understanding. We provided a new data set of temporal word analogies and tested a variant of a recently proposed temporal embedding against it. Results show a good accuracy when solving those character analogies across time. This supports that idea that these embeddings can properly understand the semantic role of each character, the results being particularly robust in case of static analogies. We also provided a visualization of the temporal embeddings to trace the evolution over time of characters in a story plot.

As a future work, we would like to use those embeddings for more CNU tasks and also moves from narratives from social media. An important application of the identification and analysis of such character-centric narratives in social media could be the identification of victims and bullies in hate-speech dialogues.

6. REFERENCES

[1] C. W. Welin, "Scripts, plans, goals and understanding, an inquiry into human knowledge structures: Roger C. Schank and Robert P. Abelson Hillsdale," *J. Pragmatics, Lawrence Erlbaum Assoc.*, vol. 3, no. 2, pp. 211–217, Apr. 1979.

[2] M. Regneri, A. Koller, and M. Pinkal, "Learning script knowledge with web experiments," in *ACL 2010 - 48th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, 2010, pp. 979–988.

[3] N. Chambers and D. Jurafsky, "Unsupervised learning of narrative event chains," in *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, 2008, pp. 789–797.

[4] M. Mark A. 1977- Finlayson, "Learning narrative structure from annotated folktales," 2012.

[5] Riedl, Mark O., and R. Michael Young. "Character-focused narrative generation for execution in virtual worlds," in *International Conference on Virtual Storytelling*, Springer, Berlin, Heidelberg, 2003, pp. 47-56.

[6] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 2013, pp. 3111–3119.

[7] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," in *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers*, 2016, vol. 3, pp. 1489–1501.

[8] R. Bamler and S. Mandt, "Dynamic word embeddings," in *34th International Conference on Machine Learning, ICML 2017*, 2017, vol. 1, pp. 607–621.

[9] T. Szymanski, "Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings," pp. 448–453.

[13] Chaturvedi, S., Peng, H. and Roth, D., 2017, September. Story comprehension for predicting what happens next. *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* , pp. 1603-1614.

[14] Vani, K. and Antonucci, A., 2019. NOVEL2GRAPH: Visual Summaries of Narrative Text Enhanced by Machine Learning. *In Text2Story@ ECIR,* pp. 29-37.

[15] Nalisnick, E.T. and Baird, H.S., 2013, August. Character-to-character sentiment analysis in shakespeare's plays. *In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,* vol.2, pp. 479-483.

[16] John, M., Lohmann, S., Koch, S., Wörner, M., and Ertl, T., Visual Analytics for Narrative Text, 2016.

[17] Labatut, V., and Bost, X., 2019. Extraction and Analysis of Fictional Character Networks: A Survey. *arXiv preprint arXiv:1907.02704*.

[18] Roemmele, M., and Gordon, A, "An Encoder-decoder Approach to Predicting Causal Relations in Stories." *Proc. of the First Workshop on Storytelling*, pp. 50-59, 2018.

[19] V. Di Carlo, F. Bianchi, and M. Palmonari, "Training Temporal Word Embeddings with a Compass," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 6326–6334, Jul. 2019.

[20] A. Kutuzov, L. Øvrelid, T. Szymanski, and E. Velldal, "Diachronic word embeddings and semantic shifts: a survey," 2018.

[21] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong, "Dynamic word embeddings for evolving semantic discovery," in *WSDM 2018 - Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018, vol. 2018-Febua, pp. 673–681.

[22] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, "Statistically significant detection of linguistic change," in *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 625–635.