Universitat Politècnica de València

Departamento de Sistemas Informáticos y Computación

Tesis de Doctorado en Informática

Mirko Lai

Language and Structure in Polarized Communities

Directores de Tesis
Giancarlo Ruffo
Università degli Studi di Torino, Italy
Paolo Rosso
Universitat Politècnica de València, Spain
Codirector de Tesis
Viviana Patti
Università degli Studi di Torino, Italy

Enero 2019

University of Turin

Doctoral School of Sciences and Innovative Technologies PhD Program in Computer Science
XXX cycle

PhD Dissertation
Mirko Lai

Language and Structure in Polarized Communities

Advisors
Giancarlo Ruffo
Università degli Studi di Torino, Italy
Paolo Rosso
Universitat Politècnica de València, Spain
Co-advisor
Viviana Patti
Università degli Studi di Torino, Italy

PhD Coordinator
Marco Grangetto

January 2019

Abstract

Policymakers and citizens intensively use social media for expressing their opinions about public debates nowadays. Furthermore, if on the one hand web interactions allow users to access to diverse viewpoints, on the other hand they do not resolve conflicts, but they often contribute to further polarize the debate. Social media platforms such as Twitter make available to researchers a huge amount of user-generated contents. To explore the public opinion and to investigate how individuals communicate with each other is now possible as never before. Interest in automatically detecting the opinions expressed in social media texts has grown significantly in the recent years. Natural language processing methods based on machine learning algorithms or deep learning approaches have been proposed for detecting users' opinion and stance towards a specific topic (person, organization, movement, policy, etc.) discussed in social media. Moreover, several works suggested that ideological segregation exists in social media despite they potentially expose users to a larger range of different point of views. In this thesis, we address the problem of stance detection in social media focusing on polarized political debates in Twitter. Stance detection consists in automatically determine whether the author of a post is in favor or against a target of interest, or whether the opinion toward the given target can not be inferred. We deal with political topics such as electoral events (e.g., political elections or referendums) and consequently the targets of interest are both politicians and referendums. We also explore the communications which take place in these polarized debates shedding some light on dynamics of communications among people having concordant or contrasting opinions, particularly focusing on observing opinions' shifting. We propose machine learning models for addressing stance detection as a classification problem. We explore features based on the textual content of the tweet, but also features based on contextual information that do no emerge directly from the text. Using the English benchmark dataset proposed for the shared tasks on stance detection held at SemEval 2016, we explore the contribution on stance detection of investigating the relations among the target of interest and the other entities involved in the debate. We particularly focus on the 2016 United States presidential primaries for the Democratic and Republican parties main candidates. Results outperform the best ones obtained by the teams participating in the task. Our model takes advantage of knowing the relations among the target of interest and the involved entities for inferring the stance even when the target is not directly mentioned. Participating to the "Stance and Gender Detection in Tweets on Catalan Independence" shared task held at IberEval 2017, we proposed other textual and contextual based features for detecting stance on Spanish and Catalan tweets. Our system (iTACOS) ranked in as the first position among ten participating teams for both languages at the stance detection sub-task. With the main aim of facing stance detection

in a multilingual perspective and having an homogeneous setting for multi-language comparisons, we collected tweets in French and Italian also. We decided to select topics which are very similar to those of the benchmarks released by SemEval 2016 and IberEval 2017 for the purpose of making the novel datasets more comparable with them. The French dataset (E-FRA) consists in tweets about the run-off of the French presidential elections held in 2017. We gathered Italian tweets about the Italian Constitutional Referendum for creating the Italian corpus (R-ITA). The multilingual extension of our stance detection model (multiTACOS) shows that stance detection is affected by the different styles used by users for communicating stance towards target of different types (persons or referendum) more than the used language. With the aim of retrieving contextual information about the social network of Twitter's users (the shared tasks usually release only the content of the tweet leaving aside information about the tweeter), we created other two datasets, one in English and one in Italian, respectively about the Brexit (TW-BREXIT) and the Italian Constitutional referendum (ConRef-STANCE-ita). In both the case studies, we show that users tend to aggregate themselves in like-minded groups. For this reason, the model takes advantage of knowing the online social community the tweeter belongs to and outperforms the results obtained by using only features based on the content of the post. Furthermore, experiments show that users use different type of communication depending on the level of agreement with the interlocutor's opinion, i.e., friendship, retweets, and quote relations are more common among like-minded users, while replies are often used for interacting with users having different stances. Addressing stance detection in a diachronic perspective, we also observe both opinion shifting and a mitigation of the debate towards an unaligned position after the outcome of the vote. Then, we observe that accessing to a larger diversity of point of views can influence the propensity to change the personal opinion. We finally show that the usefulness of features based on a graph representation of a domain of interest is not limited to stance detection, but can be applied to different scenarios. Proposing another classification task that performs talent identification in sport, particularly focusing on the case study of table tennis, we show that networks metrics based on centrality are strong signal for talent and can be used for training a machine learning algorithm model for this task too.

Abstract

Attualmente politici e cittadini utilizzano intensamente i social media per esprimere le loro opinioni sui dibattiti pubblici. Inoltre, se da un lato le interazioni web consentono agli utenti di conoscere diversi punti di vista, dall'altro non risolvono i conflitti, ma spesso contribuiscono a polarizzare ulteriormente il dibattito. Le piattaforme di social media come Twitter mettono a disposizione dei ricercatori una grande quantità di contenuti generati dagli utenti. Mai come ora è possibile conoscere l'opinione pubblica e studiare come le persone comunicano tra loro. L'interesse nell'individuare le opinioni espresse nei testi dei social media è cresciuto significativamente negli ultimi anni. Sono stati proposti metodi di elaborazione del linguaggio naturale basati su algoritmi di apprendimento automatico o approcci di deep learning per rilevare l'opinione degli utenti su un particolare argomento discusso nei social media (persone, organizzazioni, movimenti, politici, etc.). Inoltre, diversi lavori hanno suggerito che nei social media esiste la segregazione ideologica nonostante questi strumenti possano potenzialmente esporre gli utenti a una gamma più ampia di punti di vista. In questa tesi affrontiamo il problema del rilevamento delle opinioni nei social media concentrandoci su dibattiti politici polarizzati in Twitter. Il rilevamento delle opinioni (stance detection) consiste nel determinare automaticamente se l'autore di un post è a favore o contro un target di interesse, o nel caso in cui l'opinione verso il dato target non possa essere inferita. Trattiamo argomenti politici come le elezioni politiche e i referendum e di conseguenza i target di interesse sono sia persone che referendum. Esploriamo anche le comunicazioni che hanno luogo in questi dibattiti polarizzati, facendo luce sulle dinamiche comunicative tra persone che hanno opinioni concordanti o contrastanti, in particolare concentrandoci sull'osservazione dello spostamento delle opinioni (opinion shifting). Proponiamo modelli di apprendimento automatico per la stance detection affrontandolo come un problema di classificazione binaria. Esploriamo feature basate sul contenuto testuale del tweet, ma anche feature basate su informazioni contestuali che non emergono direttamente dal testo. Utilizzando il set di dati di riferimento proposto per uno shared task sulla stance detection tenutosi a SemEval 2016, esploriamo il contributo che lo studio delle relazioni tra il target di interesse e le altre entità coinvolte nel dibattito fornisce alla stance detection. In particolare abbiamo considerato i due principali candidati in corsa per le primarie del partito democratico e repubblicano in vista delle presidenziali degli Stati Uniti del 2016. I risultati superano quelli ottenuti dai team che hanno partecipato al task. Il nostro modello si avvale della conoscenza delle relazioni tra il target di interesse e le entità citate per inferire la stance anche quando il dato target non è direttamente menzionato. Participando allo shared task "Stance and Gender Detection in Tweets on Catalan Independence" svoltosi a IberEva 2017, abbiamo proposto altre feature testuali e contestuali per la stance detection sui

tweet spagnoli e catalani. Il nostro sistema (iTACOS) si è classificato primo tra i dieci team partecipanti per entrambe le lingue nel sub-task di stance detection. Affrontando la stance detection in una prospettiva multilingue, abbiamo raccolto tweet in francese e italiano. Abbiamo deciso di selezionare argomenti che sono molto simili a quelli dei dataset di benchmark pubblicati a SemEval 2016 e IberEval 2017 allo scopo di rendere i nuovi dataset più confrontabili con quelli già esistenti. Il corpus francese (E-FRA) consiste in tweet sul ballottaggio per le elezioni presidenziali francesi del 2017. Abbiamo raccolto tweet italiani sul referendum costituzionale italiano per la creazione del corpus (R-ITA). L'estensione multilingue del nostro modello (multiTACOS) mostra che la stance detection è influenzata dai diversi stili usati dagli utenti per comunicare la stance nei confronti di target di diversi tipi (persone o referendum) piuttosto che dalla lingua utilizzata. Con l'obiettivo di recuperare informazioni contestuali sulla rete sociale degli utenti di Twitter (solitamente i shared task rilasciano solo il contenuto del tweet lasciando da parte le informazioni sul twittatore), abbiamo creato altri due set di dati, uno in inglese e uno in italiano, rispettivamente sul Brexit (TW-BREXIT) e sul referendum costituzionale italiano (ConRef-STANCE-ita). In entrambi i casi di studi, mostriamo che gli utenti tendono ad aggregarsi in gruppi aventi idee simili. Per questo motivo, il modello sfrutta la conoscenza della comunità sociale online di cui il twittatore fa parte e supera i risultati ottenuti utilizzando solo le feature basate sul contenuto del post. Inoltre, le prove dimostrano che gli utenti usano diversi tipi di comunicazione a seconda del livello di accordo con l'opinione dell'interlocutore, ad esempio relazioni di friendship, retweet e quote sono più comuni tra gli utenti affini, mentre le reply sono spesso utilizzate per interagire con utenti che hanno posizioni diverse. Affrontando la stance detection in una prospettiva diacronica, abbiamo anche osservato sia l'opinion shifting, che la tendenza del dibattito a convergere verso posizioni neutre dopo l'esito del voto. Inoltre, abbiamo osservato che avere contatti con una più ampia varietà di opinioni può influenzare la propensione a cambiare la propria opinione. Alla fine, dimostriamo che le feature basate su una rappresentazione di un problema tramite grafo non si limitano alla stance detection, ma possono essere applicate ad altri diversi scenari. Proponendo un altro task di classificazione che ha come obiettivo l'identificazione del talento nello sport, in particolare focalizzandoci sul caso di studio del tennis da tavolo, abbiamo dimostrato che le metriche di rete basate sulla centralità sono un forte segnale per il talento e possono essere utilizzate per addestrare un modello basato su algoritmo di apprendimento automatico per affrontare questo compito.

Resumen

Los políticos y los ciudadanos están utilizando las redes sociales de forma intensiva para expresar sus opiniones sobre los debates públicos. Si bien por una parte las interacciones en la web permiten a los usuarios conocer diferentes puntos de vista, por otra parte no resuelven conflictos, por el contrario, a menudo contribuyen a polarizar aún más el debate. Las plataformas de redes sociales como Twitter proporcionan a los investigadores una gran cantidad de contenido generado por los usuarios. Con lo que explorar la opinión pública e investigar cómo las personas se comunican entre sí es ahora posible como nunca antes. El interés por identificar las opiniones expresadas en los textos generados en las redes sociales ha aumentado significativamente en los últimos años. Se han propuesto métodos de procesamiento de lenguaje natural basados en algoritmos de aprendizaje automático o enfoques de aprendizaje profundo para detectar la opinión de los usuarios sobre un tema específico discutido en las redes sociales. Además, varios trabajos han sugerido que en las redes sociales existe la segregación ideológica, aunque estas herramientas pueden potencialmente exponer a los usuarios a una gama más amplia de puntos de vista diferentes. En esta tesis abordamos el problema de la detección de las opiniones en las redes sociales, centrándose en los debates políticos polarizados en Twitter. La detección de opiniones (stance detection) consiste en determinar automáticamente si el autor de una publicación está a favor o en contra de un objetivo de interés, o si no se puede inferir la opinión. Nos ocupamos de temas políticos como las elecciones políticas y los referendos y, como resultado, los objetivos son tanto personas como referendos. También exploramos las comunicaciones que tienen lugar en estos debates polarizados, arrojando luz sobre las dinámicas de comunicación entre personas que tienen opiniones en acuerdo o en conflicto, enfocándonos en particular en la observación del cambio de opiniones (opinion shifting). Proponemos modelos de aprendizaje automático para la stance detection como si fuera un problema de clasificación binaria. Exploramos características basadas en el contenido del texto del tweet, además usamos características basadas en información contextual que no emerge directamente del texto. Utilizando el corpus de benchmark propuesto para la tarea compartida sobre la stance detection realizado para SemEval 2016, exploramos la contribución que el estudio de las relaciones entre el objetivo de interés y las otras entidades involucradas en el debate proporciona a la stance detection. En particular, consideramos a los dos candidatos principales que competían para las elecciones primarias del Partido Demócrata y Republicano antes de las elecciones presidenciales de los Estados Unidos de 2016. Los resultados superan los obtenidos por los equipos que participaron en la tarea. Nuestro modelo hace uso del conocimiento de las relaciones entre el objetivo de interés y las entidades mencionadas para inferir la opinión, incluso cuando el objetivo no es mencionado directamente. Al participar en la tarea "Stance and Gender

Detection in Tweets on Catalan Independence" organizado para IberEval 2017, hemos propuesto otras características textuales y contextuales para la stance detection en tweets en español y en catalán. Nuestro sistema (iTACOS) consiguió la primera posición entre los diez equipos participantes para ambos idiomas en la subtarea de stance detection. Explorando la stance detection desde una perspectiva multilingüe, hemos creado un corpus de tweets en francés y uno en italiano. Hemos decidido seleccionar temas que sean muy similares a los dos corpus de benchmark publicados en SemEval 2016 e IberEval 2017 para que los nuevos conjuntos de datos sean más comparables a los ya existentes. El corpus francés (E-FRA) consta de tweets sobre la segunda vuelta de las elecciones presidenciales francesas del 2017. Para la creación del corpus (R-ITA) hemos recogido tweets italianos sobre el referéndum constitucional italiano. La extensión multilingüe de nuestro modelo (multiTACOS) muestra que la stance detection está influenciada más por los diferentes estilos utilizados por los usuarios para comunicar la opinión sobre objetivos de diferentes tipos (personas o referendos) en lugar del idioma utilizado. Con el objetivo de recuperar información contextual sobre la red social de los usuarios de Twitter (generalmente las tareas compartidas solo consisten en el contenido del tweet, dejando de lado la información sobre el usuario), hemos creado otros dos conjuntos de datos, uno en inglés y uno en italiano, respectivamente, sobre el Brexit (TW-BREXIT) y sobre el referéndum constitucional italiano (ConRef-STANCE-ita). En ambos casos de estudio, mostramos que los usuarios tienden a agruparse en grupos con ideas similares. Por este motivo, el modelo que explota el conocimiento de la comunidad social a la que el autor del tweet pertenece, supera los resultados obtenidos utilizando solo las funciones basadas en el contenido de la publicación. Además, la evidencia muestra que los usuarios utilizan diferentes tipos de comunicación según el nivel de acuerdo con la opinión del interlocutor, por ejemplo, las relaciones de amistad, los retweets y las citas (quote) son más comunes entre los usuarios relacionados, mientras que las respuestas (replies) se utilizan a menudo para interactuar con usuarios que tienen diferentes posiciones. Al abordar la stance detection desde una perspectiva diacrónica, también observamos tanto el cambio de opinión como la mitigación del debate hacia posiciones neutrales después del resultado de la votación. Además, hemos observado que tener contacto con una variedad más amplia de opiniones puede influir en la propensión a cambiar de opinión. Finalmente, mostramos que las características basadas en una representación gráfica de un dominio de interés no se limitan a la stance detection, sino que se puede aplicar a diferentes escenarios. Al proponer otra tarea de clasificación que realiza la identificación del talento en el deporte, especialmente en el estudio de caso del tenis de mesa, mostramos que las métricas de redes basadas en la centralidad son una señal fuerte para el talento y pueden usarse para entrenar un modelo de algoritmo de aprendizaje automático para enfrentar esta tarea.

Resum

Actualment, els polítics i els ciutadans utilitzen de manera freqüent els mitjans de comunicació social per expressar les seves opinions sobre els debats públics. Si bé, d'una banda, les interaccions web permeten als usuaris accedir a diferents punts de vista, d'altra banda, no resolen conflictes, sinó que sovint contribueixen a polaritzar encara més el debat. Les plataformes de mitjans de comunicació social com Twitter posen a disposició dels investigadors una gran quantitat de continguts generats pels usuaris. Explorar l'opinió pública i investigar com les persones es comuniquen entre si ara és possible com mai abans ho havia estat. L'interès per detectar automàticament les opinions expressades en els textos de les xarxes socials ha augmentat significativament en els darrers anys. S'han proposat mètodes de processament del llenguatge natural basats en algorismes d'aprenentatge automàtic o en enfocaments de deep learning per detectar l'opinió dels usuaris i la seva posició envers un tema específic (persona, organització, moviment, política, etc.) tractats o debatuts en els mitjans de comunicació social. D'altra banda, diversos treballs suggereixen que la segregació ideològica també existeix en els mitjans de comunicació social, tot i que potencialment exposen als usuaris a una major varietat de punts de vista. En aquesta tesi doctoral abordem el problema de la detecció de posició (stance detection) en els mitjans de comunicació social, especialment centrat en els debats polítics polaritzats a Twitter. La stance detection consisteix a determinar automàticament si l'autor d'una publicació està a favor o en contra d'un objectiu o tema d'interès, o si l'opinió envers d'aquest objectiu o tema determinat no es pot inferir. Ens ocupem de temes polítics com ara esdeveniments electorals (com per exemple, eleccions polítiques o referèndums) i, en conseqüència, els temes d'interès són, en concret, la stance detection en vers dirigents polítics i referèndums. També explorem les comunicacions que es duen a terme en aquests debats polaritzats, que posen de manifest la dinàmica de les comunicacions entre les persones que tenen opinions concordants o contrastades, especialment centrant-nos en l'observació del canvi de les opinions. Proposem models d'aprenentatge automàtic per abordar la stance detection com un problema de classificació. Explorem les funcions basades en el contingut textual del tweet, però també les funcions basades en la informació contextual que no afloren directament del text. Utilitzem el conjunt de dades de referència en anglès proposat per a les tasques compartides sobre stance detection celebrades a SemEval 2016, per explorar la contribució a la stance detection d'investigar les relacions entre l'objectiu d'interès i les altres entitats implicades en el debat. Ens centrem específicament en les primàries presidencials de 2016 dels Estats Units per als candidats principals dels partits demòcrates i republicans. Els resultats que obtenim superen els millors resultats obtinguts pels equips participants en la tasca. El nostre model aprofita el coneixement de les relacions entre l'objectiu d'interès i les entitats implicades per inferir la posició, fins

i tot quan l'objectiu no es menciona directament en el tweet. En la participació a la tasca compartida de "Stance and Gender Detection in Tweets on Catalan Independence" celebrada a IberEval 2017, es van proposar altres trets textuals i contextuals per detectar la posició dels autors dels tweets, escrits en espanyol i en català, envers la independència de Catalunya . El nostre sistema, iTACOS, va quedar en primera posició entre els deu equips participants en la subtasca de stance detection tant en català com en castellà. Amb l'objectiu principal d'abordar la stance detection des d'una perspectiva multilingüe i disposar d'un entorn homogeni per a les comparacions multilingües, també hem recopilat tweets en francès i italià. Hem seleccionat temes molt semblants als utilitzats en les tasques de SemEval 2016 i IberEval 2017 per tal que aquests nous conjunts de dades siguin més comparables amb les dades de les tasques mencionades. El conjunt de dades en francès (E-FRA) consisteix en tweets sobre la segona volta de les eleccions presidencials franceses celebrades el 2017. Es van recopilar tweets en italià sobre el Referèndum Constitucional Italià per crear el corpus italià (R-ITA). L'extensió multilingüe del model de detecció de posició (multiTACOS) mostra que la detecció de posició es veu afectada pels diferents estils que utilitzen els usuaris per comunicar la posició envers objectius de diferents tipus (persones o referèndum) més que la llengua utilitzada. Amb l'objectiu de recuperar informació contextual sobre la xarxa social dels usuaris de Twitter (les tasques compartides solen publicar només el contingut del tweet i deixen de banda, en canvi, la informació sobre la persona que escriu el tweet), vam crear dos conjunts més de dades, un en anglès i un en italià, el corpus Brexit (TW-BREXIT) i el corpus del referèndum constitucional italià (ConRef-STANCE-ita) respectivament. En els dos casos, demostrem que els usuaris tendeixen a agrupar-se en grups d'opinió o creences similars. Per aquest motiu, el model aprofita el coneixement de la comunitat social en línia al qual pertany el tweeter i supera els resultats obtinguts utilitzant només funcions basades en el contingut de la publicació. És més, els experiments també mostren que els usuaris fan servir diferents tipus de comunicació en funció del nivell d'acord amb l'opinió del seu interlocutor, és a dir, les relacions d'amistat (friendship), retweets i cotitzacions (quotes) són més freqüents entre els usuaris amb idees afins, mentre que les respostes (replies) s'utilitzen sovint per interactuar amb els usuaris que tenen posicions o opinions diferents.

A l'hora d'abordar la stance detection des d'una perspectiva diacrònica, també observem el canvi d'opinió i la mitigació del debat cap a una posició no alineament després del resultat de la votació. A continuació, observem que l'accés a una major diversitat de punts de vista pot influir en la propensió a canviar l'opinió personal. Finalment, mostrem que la utilitat de les funcions basades en una representació gràfica d'un domini d'interès no es limita a la stance detection, sinó que es pot aplicar a diferents escenaris. Proposar una altra tasca de classificació que realitzi la identificació de talent en l'esport, especialment centrada en l'estudi de cas del tennis de taula, mostrem que les

xarxes mètriques basades en la centralitat són un fort senyal per a detectar el talent i també es pot utilitzar per a l'entrenament d'un model d'algorisme d'aprenentatge automàtic per a aquesta tasca.

This thesis has been revised and positively evaluated, considering it admissible for the defense, by **Rocio Abascal-Mena** (Universidad Autónoma Metropolitana Unidad Cuajimalpa, Mexico City, Mexico), **Leo Ferres** (Institute of Data Science, Faculty of Engineering, Universidad del Desarrollow & Telefónica I+D, Santiago, Chile), and **Jahnas Otterbacher**, (Open University of Cyprus, Nicosia, Cypro).

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Nowadays, social media are gaining a very important role in public debates and a significant part of the population is exposed to information though them [14, 16]. Furthermore, political leaders use social media directly to communicate with their citizens. On the other hand, citizens take part in the discussion, by supporting or criticizing their political opinions. For these reasons, social media provide a powerful experimental tool to deduce the mood of the public opinion and investigate how individuals are exposed to diverse viewpoints. The large amount of users' generated data motivated the need for new automated forms of textual content analysis.

In this thesis, we aim to explore the problem of automatic Stance Detection (SD), the task of identifying the opinion (against, favor, neutral, or none) towards a defined target of interest (person, organization, movement, policy, etc.) from a piece of text [60]. We focused, in particular, on political polarized debates held on Twitter such as the elections and/or referendums. We propose to address the problem of SD as a classification task, taking into account, not only the information derived from the textual content of the tweet, but also from the external knowledge of the context of the debate. To do so, we also take advantage from a characteristic of social media, i.e. the possibility to create social media relations among people registered on the platform. Thus, we propose to extract new features for SD task from the network structure particularly focusing in investigate the social media community in which each user belongs. We also show that features based on social network structure are useful not only in SD, but also in other classification tasks unrelated to Natural Language Processing such as talent prediction.

This chapter introduces some basic concepts of networks science and Natural Language Processing (NLP). We first describe social networks focusing on features such as communities detection and segregation. Therefore, we introduce Natural Language Processing, focusing on the task of stance detection, thus providing a window on the state of the art. Finally, we present works joining complex networks and Natural Language Processing

approaches. In the last section, we present the research questions and the contributions we produced during my Ph.D.

1.1 Network Science

Networks have been studied for so long that we need to go back to the 18st century to find the first study about networks structure. The driving force behind this inquiry was the need to solve the Seven Bridges of Königsberg problem showed in Figure 1.1. The well-known enigma consisted in verifying if there exists at least one path that crosses every bridge of the Eastern Prussian city of Königsberg (now Kaliningrad, Russia) one and only one time [32].

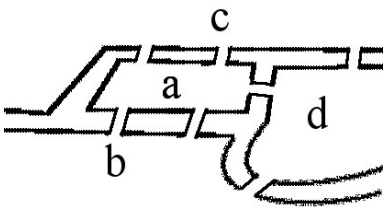

Figure 1.1: A stylistic view of Königsberg in 18st-century showing the Seven Bridges over the River Pregel connecting the four sides of the city.

In 1736, Leonhard Euler described the problem in abstract terms using the concepts of vertices (sides of the city) ad edges (bridges) for the first time as showed in Figure 1.2.

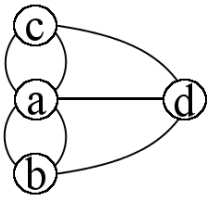

Figure 1.2: The abstract graph representation of the Seven Bridges of Königsberg problem where every side is represented by a vertex ad every bridge by an edge.

Euler showed that a complex system can be simplified representing it as a graph[1] mathematically demonstrating that a walk crossing every edge once exists only whether every vertex has an even number of edges [84]. This theorem could be considered as the foundation of graph theory, the branch of mathematics that studies the structures used to represent relations among

[1]Euler didn't call such mathematical abstraction as "graph". The word "graph" was first used in 1878 by Sylverster [90].

objects. A graph representation of a network offers a common language to study systems that belong to very different domains to such an extent that graph theory has been applied to many problems in mathematics, biology, physics, computer science, and other scientific and not-scientific areas since then [32].

In the 21st century, the advent of the Internet, a constant increase of computational resources, and the availability of graphs tools, allowed researchers to collect, organize in form of graphs, and analyze data belonging to several different complex systems. A new discipline, focusing on empirical data in place of mathematical abstract structures, arises at this juncture: *network science*. Although they all could be considered as synonyms, network scientists call graphs *networks*, vertices *nodes*, and edges *links*. Furthermore, researchers discovered that the architecture of networks are similar and governed by the same laws and principles in various domains. Subsequently, a common set of mathematical tools could be used to explore these systems belonging to different fields. This universality offers the foundation of the new discipline of network science [7, 20].

1.1.1 Fundamentals of Graph Theory

Usually a network is represented by a graph $G = (V, E)$ where V is the set of nodes (or vertices) and E the set of links (or edges). A useful way to mathematically represent G is via the adjacent matrix A where the value A_{ij} indicates the number of the links that exist between the nodes i and j where $i, j \in V$.

`Weighted and Unweighted graphs.` In a *unweighted* graph, the value of A_{ij} is 1 if at least a link exists between i and j, 0 otherwise. A_{ij} could assume a value $w_{i,j}$ that represents the weight of the link (for example, the number of kilometres between two cities or the number of calls carried out between two persons) in a *weighted* graph.

`Directed and Undirected graphs.` When we are interested in representing direct relations among entities, we use a *directed* graph. In this case, $A_{ij} \neq A_{ji} \quad \forall i, j \in V$. For example, asynchronous relations could exist in Twitter social media following based-network due to the user i could follow the user j and j is not required to follow i. Differently, Facebook social friendship based-network could be represented as a *undirected* graph due to a mutual relation of friendship can only exist. In this case, $A_{ij} = A_{ji} \quad \forall i, j \in V$.

`Degree`. One of the most important properties of each node is its *degree* k. k_i represents the number of links the node i has to other nodes. The links could be incoming or outgoing in a directed graph, accordingly, we can define in-degree k_{in}, and out-degree k_{out}. If N is the number of the nodes of G, the total number of links L of G is:

$$L = \sum_{i=1}^{N} k_i \tag{1.1}$$

The network representation with the adjacent matrix is very usefully for many tasks due to the simplicity to mathematically define a lot of metrics. For example, it is possible to define the total number of links L of G with

$$L = \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} A_{ij} \tag{1.2}$$

1.1.2 Online Social Network

Networks science has been applied in several fields including physics, brain science, transportation, power grid, biology, computer science and so on. This is because graphs can be very useful for representing either symmetric or asymmetric relations among objects, the same applies to social networks. In fact, social actors can be easily represented as nodes of a network and the different types of social ties (acquaintance, friendship, family, professional etc.) can be represented as the links that connect the nodes. Analyzing social networks needs the cooperation of many disciplines in addition to network science such as sociology, social psychology, and statistics. The interdisciplinary attention of network science could be verified by the fact that one of the most cited social network papers were wrote by a sociologist, Granovetter [31]. Another example of a great attention on social science papers relates a work written in 1954 by two sociologists, Lazarsfeld and Merton. They introduced in modern sociology the principle of *Homophily*, that is the principle that affirms that people tend to be similar to their friends [47]. Since then, several network scientists observed that social empirical data confirm the homophily principle and that people tend to create an high amount of bonds with similar and many fewer links with the other individuals [55].

The grow of Internet and the introduction of new technologies such as blogs, forums, social media, and social networking service, allows people to create new forms of social relations characterized by computer-mediated interactions: the social media relations. The physical distance is not relevant as before. While two individuals that live in the same building may not know each other, a social media could allow that a relation between two individuals sitting in front of their computers on the other side of the world exists. Furthermore, the huge amount of users generating data simultaneously allows to more easily observe social phenomena in a wide variety of disciplines compared with traditional survey data [48].

For example, the pervasive use of new technologies allows researchers to track human mobility by their mobile phone and understand individual human mobility patterns. Epidemic prevention to emergency response, urban

planning and human mobility patterns modelling could be based on these results [28].

Goncalves et al. [29] tested and validated on Twitter the theoretical cognitive limit on the number of stable social relationships a human is able to handle known as Dunbar's number.

Weng et al. [96] similarly found empirical support for the hypotheses of Granovetter's weak tie theory of social networks [31] in Twitter. They showed that strong social media ties (i.e., steady relations such as friendship, in contrast with weak ties that include sporadic social ties formed among acquaintances) carry the large majority of interaction events. Furthermore, the authors observed that *attention* is high in weak ties and in very strong ties. Sticking with Twitter platform, Conover et. al [16] showed that users tend to retweet posts supporting the same political orientations.

Exploring an independent information and communication platform for Swiss politics, Garcia et. al [24] measured network polarization among politicians exploring both the relation between ideology and social structures in online interactions.

Adamic et. al [1] observed that blogs preferentially link to other blogs of the same political ideology and that the "value" homophily is involved as theorized by Lazarsfeld and Merton [47] that observed that people tend to bond with others who think in similar ways, regardless of any differences in their status characteristics (i.e. gender, age, social status).

1.1.3 Community Detection

A large section of network science literature deal with communities. In many different fields we can define communities, whose structure definition do not change [80, 6]. Although, we only focus on social media network here.

Researchers studied social network for many years with the aim of identifying like-minded individuals [102], therefore they started to use community detection for identifying sub-groups of nodes with latent common features using only the connectivity structure of the network People tend to create relations with friends, schoolmates, and coworkers etc. Consequently, groups consisting in circles of friends, or in individuals who go to the same school, or work in the same company could be observed in the social network.

`Defining Communities`. Communities could be therefore defined as *locally dense connected subgraphs in a network* [7, Chapter 9]. Over the years several definitions of what *a dense connected subgraph* is have been proposed:

- *Cliques*. Each node of the community C is connect with all the other nodes belonging to the community.

$$A_{ij} \neq 0 \quad \forall i, j \in C \tag{1.3}$$

- *Strong community.* Each node of the community has more links to other nodes of the same community then to nodes belonging to other communities.

$$\sum_{j \in C} A_{ij} > \sum_{j \notin C} A_{ij} \quad \forall i \in C \tag{1.4}$$

- *Weak community.* The sum of internal links within the community exceeds the sum of external links among other communities.

$$\sum_{i,j \in C} A_{ij} > \sum_{i \in C j \notin C} A_{ij} \tag{1.5}$$

`Graph Partitioning`. The simple way to partition a graph into communities is called *graph bisection*. The process consists in detecting two not overlapping communities of the same size minimizing the number of links (*cut off*) between the nodes of the two communities C_i and C_j and inspecting all possible partitions $p \in P$ of the graph G:

$$p = \min_{p \in P} \frac{\sum\limits_{i,j} A_{ij} \quad [C_i \neq C_j]}{\sum\limits_{i,j} A_{ij} \quad [C_i = C_j]} \tag{1.6}$$

The number of partitions consisting in two subsets of $\frac{N}{2}$ elements, disregarding their order, from a set of N elements is given by:

$$\binom{N}{\frac{N}{2}} = \frac{N!}{\frac{N}{2}!\frac{N}{2}!} \tag{1.7}$$

The complexity exponentially increases with the number of nodes of the graph. Another disadvantage is that the number and the size of the communities must be specified in advance.

`Community Detection`. Community detection faces the problem of partitioning a graph when both the number and the size of the communities are unknown in advance. The complexity with a brute-force strategy grows faster than exponentially with the increasing of N. The number of possible partitions of N elements in k groups is given by the Bell number B_N:

$$B_{N+1} = \sum_{k=0}^{N} \binom{N}{k} B_k \tag{1.8}$$

Given the impossibility to inspect all partitions of a large network, more efficient methods of partitioning a graph are necessary, therefore many approaches have been proposed. They identify communities not inspecting all possible partitions of a graph, but focusing on the ones that best satisfy one of the definitions of community as we reported earlier.

`Louvain Algorithm`.

In this thesis we have used the *Louvain Algorithm* [9] when we aimed to perform graph partitioning. This method is optimized for large networks and its complexity is in the order of $O(N \log N)$. The Louvain Algorithm is based on *modularity* Q, a metric that compares the density of links within the communities with the density of links among communities. The value of the modularity Q=0 should represent a network with a number of within-community edges not higher than a null model, where connections are placed randomly, and the formation of a community-based structure is unlikely. Whereas, the value is positive when the observed number of edges within groups exceeds the number of expected edges within groups in a random graph, negative otherwise. Defining m as the number of links and the quantity δ_{ij} as 1 if users i and j belong to the same community and 0 otherwise, we can then express the modularity as:

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij} \tag{1.9}$$

Louvain Algorithm iteratively assigns a community to each node with the aim to maximize the modularity. The method consists in two steps:

- The first step consists in assigning a different community c for each node $i \in V$. Then, the algoritm evaluates the gain in modularity Q if the node i moved into an other community. Finally, i is assigned to the community with the largest gain in modularity. This process applies for each $i \in V$.

- The second step consists in creating a new network merging the nodes belonging to the same community into a single node. The number of nodes consequently decreases at each step. The weight of the link between two aggregated groups of nodes is the sum of the weights of the links between the nodes in the corresponding communities.

The two steps are repeated until a new reassignment of communities does not further increase the modularity.

1.1.4 Polarization and Segregation

`Polarization`. Sunstein [89] discussed the phenomenon of group polarization in social relations drawing the attention of its implications for law and political theory. He affirmed that two people, who only slightly disagree with each other, will tend to be even more opposed, after they have talked to each other. This phenomenon could also explain the emergence of extreme and radical tendencies in social media communities. Polarization could also

be explained in therms of *homophily* due to people tend to bond with like-minded groups and, developing extreme views, they lead to cut relations with opposed like-minded individuals.

`Segregation`. Some studies suggested that segregation exists in social media despite they potentially expose users to a larger range of different views [21, 92]. These are mainly due to the existence of *echo chambers* and *filter bubbles* where individuals hardly get exposed to information from other groups.

We introduced the previous concepts of network science with the aim of inspecting social media community and observing polarization and segregation in communications among Twitter's users about political debate. In the following section we will introduced some concepts of Natural Language Processing in order to investigate not only the structure of the graph based on users' relation, but also the textual content of the users' tweets.

1.2 Natural Languages Processing

As network science, natural languages processing (NLP) is an interdisciplinary field. Indeed, linguistics, computer science, artificial intelligence, and mathematics are involved with the aim of using computers for understanding and manipulating natural language. In general, NLP is the scientific study that concerns with the interactions between computers and human natural languages. Many problems within NLP apply to large categories of tasks for both speech and text. Nowadays, several NLP tasks are used every day on the Internet and on mobile applications such as spam filtering, recommendation in search, assisting chat bots, speech recognition, machine translation, and many others. Actually machine translation was one of the first investigated applications for NLP. The Georgetown experiment, the first public demonstration of machine translation, was done in 1954 and consisted in a fully automatic translation of about sixty Russian sentences into English [37]. The experiment created big expectations for solving the machine translation problem in few years and brought the research community attention to computational linguistics. NLP has also been involved in many other tasks such as parsing, summarization, duplicate detection, part of speech tagging, name entity recognition, text classification, and many others.

1.2.1 Automatic Text Classification

Text classification is a task that consists in assigning a text to one or more classes or categories. Many are the applications in which text classification is involved such as spam detection, sentiment analysis or language identification. Spam detection aims to automatically predict the class of a text between two classes: *regular mail* and *spam mail*. Sentiment analysis is also a binary classification task. It aims to automatically predict the polarity

of a text between two classes: *positive* and *negative*. A ternary classification is performed where the class *neutral* is considered. Another example of multiclass classification is language identification where the task consists in identifying the language of a text among any of the set of considered languages.

In order to propose and evaluate a classification method, a *corpus* of labeled texts is required. A corpus is a collection of documents, as utterance or sentences, assumed to be representative of and usable for lexical, grammatical, or other linguistic analysis [27]. Usually, a group of human annotators labels each document assigning one of the possible classes to it. Annotation schema and guidelines supply the meaning of each label for guiding the annotation process. The manual annotated process generates the so called *gold standard*: a collection of labeled documents where each label is accepted as the most valid one.

Every automatic text classification method must be evaluated by comparing its prediction with a gold standard. In a *supervised framework*, the gold standard is first divided into two subsets: *training* and *test*. The training set is used for training the model and the test set is used for comparing the predicted label with the gold label. Several methods for dividing the corpus in training and test set exists. One of the simplest and most common spitting methods consists in randomly dividing the corpus in 80% of training documents and in 20% of test ones. A common method is also to perform a k-fold validation (usually 5-fold or 10-fold) dividing the corpus in k folds and using each fold once as test and the remaining $k-1$ folds as training. Differently, no training set is available in a *unsupervised framework*.

The performance can by evaluated using several metrics. Two measures are commonly used for evaluating a classification method: *accuracy* and F_1-score. *Accuracy* (ACC) evaluates the number of correct predictions (*true positive*) divided by the total number of predictions (*predictions*), multiplied by 100 to turn it into a percentage.

$$ACC = \frac{true\ positive}{predictions} 100 \tag{1.10}$$

Accuracy could be not enough, especially in the case in which strongly unbalanced label distribution exists. If the model predicts all the documents with the most common label, the value of the accuracy is equal to the more common label frequency. A more accurate measure that takes into account both the precision p and the recall r is F_1-score.

Let p be the number of correct true predictions divided by the number of all true predictions (*precision*) returned by the classifier and r the number of correct positive results divided by the number of all positive documents (*recall*), as showed in Figure 1.3. F_1-score is calculated as:

$$\mathrm{F}_1\text{-score} = 2\frac{p \cdot r}{p + r} \tag{1.11}$$

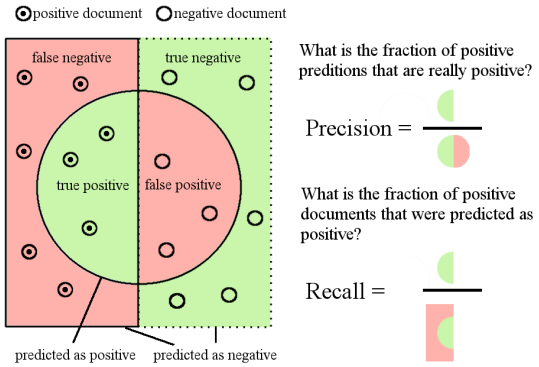

Figure 1.3: Precision and recall representation. The color green represents the correct predictions, the red the wrong ones. The documents within the circle were predicted as true, the other ones were predicted as negative.

We can measure the F-macro as the average among the F_i-score of each class by:

$$\text{F-macro} = \frac{1}{N}\sum_{i=1}^{N} \text{F}_\text{i}\text{-score} \tag{1.12}$$

1.2.2 Main Approaches to Text Classification

Text classification could be performed choosing among several types of approaches. Regular expressions or context free grammars are often used in NLP, for example, for part of speech tagging or named entity recognition. Defining a set of rules for classifying texts could be very accurate when rules are written by experts and the classification criteria can be easily controlled due to a small number of defined rules. Also automatic rules generators are exploited, but they are rarely used for text classification, and neural networks and probabilistic modeling are generally preferred.

Neural networks (or deep neural networks if more then one hidden layer is used) are employed in artificial intelligence in several tasks as image recognition. Neural networks were recently strongly employed in NLP tasks. Several works propose neural networks approaches for improving the state of the art also in text classification, but these methods often achieved very similar results compared to traditional machine learning algorithms based on probabilistic models.

Probabilistic modeling and traditional machine learning algorithms are strongly employed in text classification tasks. The availability of several machine learning algorithms such as logistic regression, naive Bayes classifi-

cation, decision trees, and support vector machines allow researchers to find the most suitable model to a wide variety of tasks. If on the one hand, researchers must not underestimate the importance of choosing the most suitable algorithm for addressing a text classification task, on the other hand, the choice of the features is essential.

In machine learning, numeric features are usually used for representing a text with a feature vector. Binary *Bag of Word* is a common feature representation: a text is represented by a vector where each index corresponds to a word, the values is 1 if the word is present in the text, 0 otherwise. Binary variables could be replaced by another numerical statistic representation as term frequency-inverse document frequency (tf-idf) that reflects the importance of a word in the corpus. Tf–idf could contribute to give few importance to words with a too low (i.e., rare word, typos) or too high (i.e. stop word) frequency.

Bag of Word is also called 1-gram feature due to the vector space includes only single words. Also 2, 3-grams are usually exploited in text classification tasks and they consist in representing the text with a vector space that contains couples or triplets of consecutive words.

Text classification tasks as language identification take advantage of char n-grams [74]. These features consider n consecutive chars as token (rather than words) for the vector representation of the text.

1.2.3 Stance Detection

Sentiment analysis and opinion mining are often exploited to monitor people's mood extracting information from users' generated contents in social media [70]. Recent trends focus on a new text classification task: stance detection [3, 50].

Several works deal with the automatic classification of movies or product reviews as positive or negative [70, 36]. It was also addressed the issue of identifying the appraisal of the quality of some aspects of the analyzed subject (e.g. keyboard, display, battery of a mobile phone)[4].

Stance detection consists in detecting users stance towards a particular *target* of interest (i.e. a person, a brand, an opinion) differentiating themselves from sentiment analysis. One of the first works that could be considered more related to SD tried to identify the *perspective* from which a document is written [50]. Investigating textual sentences about the Palestinian-Israeli conflict topic, authors used a Naive Bayes-based classifier in order to detect the views of the author of the post towards the topic.

Nevertheless, as far as we know, Somasundaran and Wiebe [86] were the first one to focus on detecting the stance towards a target rather than the polarity of a sentence. They presented (in a unsupervised framework) a stance recognition method for debate-side classification (i.e. recognizing which stance a person is taking) from web blogs. The method is based on

the association among preferences with opinions towards different aspects.

Anand et al. [3] tried to automatically classify both rebuttal and stance from a corpus of messages across several topics posted on ConvinceMe.net. They achieved higher values of accuracy for rebuttal classification and observed that the rebuttal posts are harder to classify. For what concerns stance detection, authors highlighted the importance of knowing of the context proposing a feature based on the parent posts. They also explore cue words, sentiment lexica (LIWC [72]), punctuation marks, syntaxes, and opinion dependencies. Using a Naive Bayes classifier, obtained results are good for some topics, but not for others. Overall, aggregating posts over all topics does not help achieving the baseline results obtained with unigrams.

The first shared task on SD in Twitter was held at SemEval 2016 Task 6 [60]. The task is described as follows: "Given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is in favor of the target, against the given target, or whether neither inference is likely". The task concerns six commonly known targets in the United States, such as: "Atheism", "Climate Change is a Real Concern", "Feminism Movement", "Hillary Clinton", "Legalization of Abortion", and "Donald Trump". Standard text classification features such as n-grams and word embedding vectors were exploited by the majority of the teams that participated in the task. Sentimental resources, such as EmoLex [63], MPQA [99], Hu&Liu [36], and NRC Hashtag [61], were also exploited. The best result was obtained by a deep learning approach based on a recurrent neural network trained with embeddings of words and phrases initialized with the word2vec skip-gram medel [103].

Machine learning algorithms and deep learning approaches were also exploited in a second shared task on stance and gender detection in Twitter held at IberEval 2017 [91]. The dataset concerns the political debate about *Independence of Catalonia* during the Catalan regional election that was held on September 2015. With regard to SD, participating teams employed different kinds of features such as bag of words, bag of parts-of-speech, n-grams, word length, number of words, number of hashtags, number of words starting with capital letters, and so on. In this case, we obtained the best results proposing an SVM classifier that exploited three groups of features: *Stylistic* (bag of: n-grams, char-grams, part-of-speech labels, and lemmas), *Structural* (hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet), and *Contextual* (the language of each tweet and information coming from the URL in each tweet) [42].

1.3 Network Analysis for Natural Language Processing

When we want to understand a complex system we need to know how its components interact with each other. Indeed, some works shed some light on the relation between social media network structure and information extracted from posted contents.

On one hand, linguistic approaches could help to perform community detection. For example, Xu et al. [101] introduce the concept of *sentiment community*, trying to identify online communities with similar sentiment. Performing the task of community detection, the authors show that maximizing both the intra-connections of nodes and the sentiment polarities performing community detection. The ratings of movies collected from Flixster are used.

Similarly, Deitrick et al. [17] combined sentiment analysis and community detection techniques by using Twitter's relations among users and sentiment classification of tweets. Iteratively they increased the edge weights in a social networks based on follower and friend relations in order to detect communities.

The analysis of the network structure could help to deal with linguistic tasks. Indeed, West et al. [97] joined textual and social network information for predicting the polarity of the relation between two users. They show that the model improves results individually obtained by textual and social network information.

Other authors recently explore communities for predicting stance detection. For example, Fraisier et al. [69] addressed stance detection employing community detection in networks based on different *proximities* among users. Authors explore both content-based and social-based proximities for accurately predicting stance with less than 1% of annotated profiles and considering more than two possible stances. For what concerns social-based *proximities* in Twitter, authors explored users' retweets, mentions and friends list.

These preliminary results lead us to perform the SD task using the information that could be extracted from the social network structure. We show that social communities can play a crucial role in determining stance within polarized debates [46, 45]. In our experiment, analyzing two different political debates, we observe that a strong relation exists between user's stance and social media community the user belongs to.

Twitter is a social networking service and micro-blogging site where users post messages called *tweets*. For what concerns the peculiarity of this platform, Twitter allows users to create asymmetric relations; in-link relations are called *followers*, while out-link ones are called *friends*. Users can also share with their followers the tweets written by their friends, and these tweets

are known as *retweets*. Furthermore, users can add their own comments before retweeting making it a *quote*. Finally, the last option is to answer to another user's tweet, generating a so called *reply*. More replies can form complex *conversations*. Another important feature in Twitter is the possibility of indexing tweets through the embedding of an *hashtag* (a relevant and meaningful keyword preceded by the # sign) in the text.

The several number of communication typologies among users (friendships, retweets, quotes, and replies) provides us a complex representation of users' relations inside the social network itself. Additionally, the presence of *hashtags* in the textual content allows more accurate topic filtering. Moreover the platform offers very useful tools to perform the analysis, as Twitter's REST and Streaming API. Political leaders make an extensive use of this platforms to communicate with citizens, that, on the other hand, join in online discussions supporting or criticizing their political opinions. It is not surprising indeed the scientific interest for this platform whose posts were used for both the two shared task held on SD [60, 91]. For all these reasons we used Twitter for gathering data and for performing our research.

1.4 Research Questions

Our main purpose is to explore stance in political polarized debates on Twitter. First, we focus on linguistic approaches exploring if content and contextual information, in particular information based on social network, could improve SD. We also show as features based on network structure could be useful in other classification task unrelated to SD such as talent prediction.

Then we focus on computational social science approaches exploring the diachronic and spacial evolution of the debate with the aim of inspecting user behaviour on social media in polarized contexts. We focus on observing if relevant events and social media relations could influence opinion shifting. Finally, we also explore if stance could influence the type of communication that the tweeter establishes with other users. The research questions we aimed to answer in our work could be summarized as follows:

i *Is contextual information useful for SD in social media?*

We proposed and explored the contributions of different contextual information in SD from the following hypothesis:

- *The relations among the involved entities in the debate.* When the target of interest is not mentioned, stance could be inferred knowing the relation among the target and the mentioned entities. For example:

  @realDonaldTrump No more Bush or Hillary Clinton #SemST

  In this example the target of interest is "Donald Trump". When the tweeter expresses a negative opinion towards "Bush" and "Hillary

Clinton" (two opponents of the target, the first is an opponent inside the Republican party, the second one is a Democratic candidate), the annotator should infer that the tweeter expresses a positive opinion towards the target (Chapter 2).

- *The language used by the user.* For example, when the debate takes place in a region characterized by a strong bilingualism, the language chosen by the tweeter who posts the tweet could help to infer the user's stance (Chapter 3).
- *The web source (url) the tweeter shares in their posts.* Users tend to share information they agree on, therefore, sharing a news from an online newspaper rather than another one could help to detect the stance of the tweeter (Chapter 3).
- *The online social network community the tweeter belongs to.* Users tend to create relations with like-minded people. For this reason we thought that inferring the stance of a community could help to detect the stance of those tweeters which are part of it (Chapter 5).

ii *Among the contextual features, what is the impact of the network structure?* We explore what type of social network structure further helps on SD. We analyze four different types of interaction that could be established in Twitter: friendship, retweet, quote, and reply. (Chapters 5 and 6.4)

iii *How to address SD from a multilingual perspective?* We analyzed four different political debates in five different languages such as English, Spanish, Catalan, French, and Italian (Section 6.3)

iv *Could the feature based on social network structure be used in other classification tasks?* Networks are very useful for representing complex problems such as social relations. Therefore, we try to answer to this question proposing a classification task that performers talent identification in sport particularly focusing on the case of study of table tennis (Section 6.6). We aim to prove the universality of the approach know as "embedded networking" [30].

Analyzing the debate focusing on both users' stance and relation in a diachronic perspective, we finally propose the following questions (Chapters 4, 5, and Sections 6.4, and 6.5):

v *Are there benefit for addressing SD from a diachronic perspective?* Users could change their stances during the debate after relevant events happened. We propose to split the debate in different time windows following the same users during the debate for analyzing if an opinion shifting could be observed.

vi *Could stance label distribution change during time?* We hypothesize that the frequency of NONE stances could increase when the debate approaches the end.

vii *Could the stance of the tweeter influence the type of relation s/he creates with other users?* We aim to explore what type of communication more likely bind two users that share the same opinion and more likely connect two users that have different stance.

viii *Could the neighbours influence the probability of observing an opinion shifting?* We explore if users with a high number of neighbours having different stance are more likely to change their stance in the future than users mainly connected to like-minded users.

1.5 Contributions

Stance detection has been identified as a not trivial task independent from sentiment analysis. Indeed, if on the one hand, sentiment analysis aims to detect the sentiment expressed in a piece of text, on the other, stance detection seeks to identify the user's opinion toward a defined target of interest (not necessarily mentioned in the text). In this thesis we concentrated our attention on online political debated and we faced stance detection as a classification task proposing different type of features, in particular, increasingly focusing on contextual ones. Our analysis reveals a strong relation between stance and social media communities. This triggered us to deeper explore users' behaviour inspecting social media relations and the dynamic of the polarization of political debates in a temporal perspective. The achievements of our research could be summarized as following:

- We presented a brief description of the approaches proposed in the literature particularly focusing on the two shared tasks on Stance Detection held at SemEval 2016 [60] and IberEval 2017 [91]. Our method, obtaining the highest result at IberEval 2017 and amounting the state of the art achieved at Semeval 2016, validates the assumption that contextual features could be useful for the task of stance detection (Chapters 2 and 3).

- We created four new annotated corpora of tweets for stance detection: the English TW-ChronosBrexit, the Italian ConRef-STANCE-ita, and the E-FRA and R-ITA corpora respectively in French and Italian (Chapters 4 and 5, and Section 6.3)).

- Facing stance detection in a multilingual perspective, we detected linguistic characteristics peculiar of each language. Furthermore, we showed that results are affected by the different styles used by users for

communicating stance towards target entities of different types (persons or referendum) (Section 6.3).

- We observed, on two different political debates (Chapter 4 and Chapter 5), that users tend to aggregate themselves in like-minded groups. For this reason, we proposed a contextual feature based on the community the users belong for detection their stance. The results outperform those obtained by using only features based on the content of the post.
- We show how, representing a complex problem with a network, could be useful for extracting features from the network structure for dealing with other classification task suck as talent prediction (Section 6.6).
- Users use different type of communication depending on the level of agreement with the interlocutor's opinion. Friendship, retweets, and quote relations are more common among like-minded users, while replies are often used for interacting with users having different stances (Section 6.5).
- Approaching on stance detection in a diachronic perspective, we observed both opinion shifting and a mitigation of the debate towards an unaligned position after the outcome of the vote. In a deeper analysis, results tend to shows that users having heterogeneous relations tend, approaching the end of the debate, to more likely keep their opinions unclear than user having homogeneous links (Sections 6.4 and 6.5).

1.6 Structure of the Thesis

This thesis consists in a collection of our most relevant publications about the research project I was involved to during my Ph.D. Four papers have been published in the proceedings of international conferences. One of these papers describes our approach that obtained the best results in the shared task on stance detection. One paper have been published in an international journals.

A brief overview of each mentioned paper is presented below. Then, in Chapter 6 we summarize the results we obtained in the framework of our research. In Chapter 6, we also show some unpublished results obtained during the Ph.D.. Finally, we draw some conclusions and discuss future work in the final Chapter 7.

Chapter 2. Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets

This chapter contains the first result of our research on political debates in social media that investigates stance detection. The paper has been pub-

lished in the proceedings of the *15th Mexican International Conference on Artificial Intelligence*. Starting from a benchmark dataset of English tweets released at the first shared task on stance detection (SemEval-2016 Task 6), we propose a feature based on the context surrounding the targets of interest. In particular, we define the two concepts “enemies" and “friends" for denoting the possible relations among the target and the entities related to the target. Namely, we try to model that when a tweeter is against an “enemy"/“friend" of the target, then the tweeter is in favor/against the target, and vice versa. Since our particular interests in political debates, we focus on the two targets related to the political campaign for the 2016 U.S. presidential elections: Hillary Clinton and Donald Trump. Our results, that take advantage from the proposed feature, outperform the best ones obtained by the teams participating in the task. We show that the information about “enemy" and “friend" of politicians helps in detecting the stance towards them.

Chapter 3. iTACOS at IberEval2017: Detecting Stance in Catalan and Spanish Tweets

This chapter provides a technical report including a brief description of our approach, an illustration of our experiments, and an analysis of our results for our submission for the *Stance and Gender Detection in Tweets on Catalan Independence* shared task held at IberEval-2017. The released dataset consists in Catalan and Spanish tweets about the regional elections in Catalonia (Spain) held in September 2015. The election has been explained as a de facto referendum on the possible independence of Catalonia from Spain. For this reason, the organizers chose ”independence of Catalonia“ as target for the stance detecting task. Our system (iTACOS) ranked in as the first position among ten participating teams for both languages at the stance detection sub-task. Our approach, based on *context* and *structural* features, shows that contextual features helps in stance detection even when the target of interest is not a person.

Chapter 4. Extracting Graph Topological Information and Users' Opinion

This chapter contains the paper included in the proceedings of the international conference *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*. In this paper, we explore in depth opinion shifting applying the 2016 US Primary Presidential Election as case of study. We created the TW-ChronosBrexit corpus for stance detection that we used for training a model for automatically estimate the stance of all users of our dataset. We shown that users having the same stance towards this topic tend to belong to the same social network community. Moreover, we found evidences that the neighbours are more likely to have similar opinions.

Chapter 5. Stance Evolution and Twitter Interactions in an Italian Political Debate

In this chapter we present the research work published in the proceedings of the *23rd International Conference on Natural Language & Information Systems (NLDB 2018)*. We created the ConRef-STANCE-ita corpus for stance detection for inspecting stance detection at user level and in a diachronic perspective applying the 2016 referendum on the reform of the Italian Constitution as case of study. Here, we investigate in depth social network exploiting different types of relations such as retweets, quotes, and replies. The analysis shows that users with the same stance towards a particular issue tend to belong to the same social network community. For this reason, we propose three new features for stance detection based on the online social community the user belongs. The performed experiments show that the accuracy of stance detection prediction is considerably improved adding features derived from communities extracted from retweets-based and quotes-based networks to content-based ones. This does not happen using the feature based on the communities extracted from the replies-based network. Indeed, the users mainly reply to other users with a similar opinion and we observe about 20% of cross-stance edges among them. We also shed some light on users' opinion shift dynamics observing that in this debate, users tend to be less explicit on their stance as the outcome of the vote approaches.

Chapter 6. Discussion of the Results The chapter summarizes the obtained results and presents extended experiments we carried out. First, we deeply analyze our system (iTACOS) ranked in as the first position in the *Stance and Gender Detection in Tweets on Catalan Independence* shared task held at IberEval-2017. Then, we propose an extended version of iTACOS for classifying stance in a multilingual scenario (MultiTACOS). We also carry out a qualitative analysis of the features used for addressing stance detection in the debate about the BREXIT referendum, and after, we analyze the communication among users with similar and divergent viewpoints in the Italian Constitutional referendum case of study. Finally, we explore the features extracted from a network structure in a task different from stance detection e.g. talent identification in sport particularly focusing on the case of study of table tennis.

Chapter 7. Conclusion and Future Work In this chapter we draw conclusions from the results presented in this thesis. Furthermore, the chapter outlines our publications during the Ph.D. Finally, we propose some future research lines for this work.

Chapter 2

Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets

Published in:

Lai M., Hernández Farías D.I., Patti V., Rosso P. (2017) Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. In: Sidorov G., Herrera-Alcántara O. (eds) Advances in Computational Intelligence. MICAI 2016. Lecture Notes in Computer Science, vol 10061. Springer, Cham, pages 155-168.
ISBN: 978-3-319-62433-4
DOI: https://doi.org/10.1007/978-3-319-62434-1_13

Abstract

Stance detection, the task of identifying the speaker's opinion towards a particular target, has attracted the attention of researchers. This paper describes a novel approach for detecting stance in Twitter. We define a set of features in order to consider the context surrounding a target of interest with the final aim of training a model for predicting the stance towards the mentioned targets. In particular, we are interested in investigating political debates in social media. For this reason we evaluated our approach focusing on two targets of the SemEval-2016 Task 6 on Detecting stance in tweets, which are related to the political campaign for the 2016 U.S. presidential elections: Hillary Clinton vs. Donald Trump. For the sake of comparison with the state of the art, we evaluated our model against the dataset released in the SemEval-2016 Task 6 shared task competition. Our results outperform the best ones obtained by participating teams, and show that information about enemies and friends of politicians help in detecting stance towards them.

2.1 Introduction

Social media provide a way for expressing opinions about different topics. From this kind of user-generated content it is possible to discover relevant information under several perspectives. A wide range of research has been carried out in order to exploit the vast amount of data generated in social media. One of the most interesting research areas concerns to investigate how people expose their feelings, evaluations, attitudes and emotions. These kinds of aspects are the subject of interest of Sentiment Analysis (SA) [51].

Determining the subjective value of a piece of text is the most general task of SA. Recently, the interest on studying finer-grained and different facets of sentiment in texts has derived in areas such as *Aspect based sentiment analysis* [73] and *Stance Detection* (SD) [60], which is the focus of our work. Identifying the speaker's opinion towards a particular target is the main goal of SD. It is not enough to recognize whether or not a text is positive/negative/neutral but it is necessary to infer the point of view of the tweeter towards a particular target.

Stance detection could not only provide useful information for improving the performance of SA but it could also help to better understand the way in which people communicate ideas in order to highlight their point of view towards a particular target entity. This is particularly interesting when the target entity is controversial issue (e.g., political reforms [12, 87]) or a polarizing person (e.g., candidates in political elections). Therefore, detecting stance in social media could become a helpful tool for various sectors of society, such as journalism, companies and government, having politics as an especially good application domain. Several efforts have been made in order to investigate different aspects related to social media and politics [53]. We are interested in political debates in social media, particularly in

the interaction between polarized communities. We consider that being able to detect stance in user-generated content could provide useful insights to discover novel information about social network structures. Political debate texts coming from social media where people discuss their different points of view offer an attractive information source.

This year, for the first time a shared task on stance detection in tweets was organized [60]. Two of the targets considered in order to evaluate stance detection systems were: Hillary Clinton and Donald Trump[1]. Both targets have been the focus of different research, for instance in [83] the authors studied their speeches during the 2016 political campaign. In such way, studying these targets is an attracting topic of research due to the impact of the use of social media during the political campaign for the 2016 U.S. Presidential elections.

Our approach to detect stance in tweets relies mainly on the context of the targets of interest: Hillary Clinton and Donald Trump. Besides, we also took advantage of widely used features in SA.

The paper is organized as follows. Section 2 introduces the first shared task on Twitter stance detection. Section 3 describes our method to detect stance by exploiting different features. Section 4 describes the evaluation and results. Finally, Section 5 draws some conclusions.

2.2 Detecting Stance on Tweets

The SemEval-2016 Task 6: Detecting Stance in Tweets[2] was the first shared task on detecting stance from tweets. Mohammad et. al in [60] describe the task as: *Given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is in favor of the target, against the given target, or whether inference is likely.*

Let us to introduce the following example[3]:
Support #independent #BernieSanders because he's not a liar. #POTUS #libcrib #democrats #tlot #republicans #WakeUpAmerica #SemST

The target of interest is "Hillary Clinton". Here, the tweeter expresses a positive opinion towards an adversary of the target. Consequently the annotator inferred that the tweeter expresses a negative opinion towards the target. As can be noticed, this tweet does not contain any explicit clue to find the target.

For evaluating the task, the organizers annotated near to 5,000 English tweets for stance towards six commonly known targets in the United States: "Atheism", "Climate Change is a Real Concern", "Feminism Movement",

[1] They are the candidates who won the Party Presidential Primaries for the Democratic and Republican parties, respectively.

[2] http://alt.qcri.org/semeval2016/task6/

[3] This tweet was extracted from the training set of SemEval-2016 Task 6.

“Hillary Clinton", “Legalization of Abortion", and “Donald Trump" (Stance Dataset, henceforth). A set of hashtags widely used by people when tweeting about these targets was compiled; then it was used to retrieve tweets according three categories: in-favor hashtags, against hashtags and stance-ambiguous hashtags. The tweets were manually annotated by crowdsourcing. More details about the Stance Dataset can be found in [60].

The participants in the SemEval-2016 Task 6 were required to classify tweet-target pairs into exactly one of three classes: *Favor*: It can be inferred from the tweet that the tweeter supports the target (e.g., directly or indirectly by supporting someone/something, by opposing or criticizing someone/something opposed to the target, or by echoing the stance of somebody else); *Against*: It can be inferred from the tweet that the tweeter is against the target (e.g., directly or indirectly by opposing or criticizing someone/something, by supporting someone/something opposed to the target, or by echoing the stance of somebody else); and *Neither*: None of the above.

The SemEval-2016 Task 6 was divided into two subtasks:

- Task A. Supervised Framework. The participating systems were asked to perform stance detection towards the following targets: “Atheism", “Climate Change is a Real Concern", “Feminism Movement", “Hillary Clinton", and “Legalization of Abortion". For evaluation the organizer provided a training (2,914 tweets) and test (1,249 tweets) sets.

- Task B. Weakly Supervised Framework. The task was detecting stance towards one target “Donald Trump" in 707 tweets. For this task the participants were not provided with any training data about this target.

Nineteen teams participated in Task A while only nine competed in Task B. It is important to highlight that only two systems were evaluated specifically on Task B. Figure 1 shows a brief summary of the systems. Further information about the systems in the task can be found in [8][4].

Both tasks were addressed in similar ways. Most teams exploited standard text classification features such as n-grams and word embedding vectors. Besides, some SA features from well-known lexical resources, such as *EmoLex* [63], *MPQA* [99], *Hu and Liu* [36] and *NRC Hashtag* [61], were used to detect stance in tweets. Furthermore, some teams decided to take advantage of additional data by harvesting Twitter using stance-bearing hashtags in order to have more stance tweets. It is important to highlight that the best system in Task A (MITRE) did use this alternative. A similar approach was adopted by the three best ranked systems on Task B (pkudblab, LitisMind, and INF-UFRGS). For what concerns to Task B, in order to deal with the lack of training data, some systems attempted to generalize the supervised

[4]Notice that not all the reports describing systems and approaches of teams participating at SemEval-2016 Task 6 are available in [8].

data from task A in different ways such as defining rules or by exploiting multi-stage classifiers.

Table 2.1: Brief description of the participating systems at SemEval-2016 Task 6

System	Description
MITRE [103] Task A	**Overall approach:** Recurrent neural networks. **External resources:** Words embeddings with the word2vect skip-gram method. Near to 300,000 tweets containing hashtags related to the targets.
pkudblab [95] Tasks A and B	**Overall approach:** Convolutional neural network. **External resources:** Words embeddings using the Google News dataset.
TakeLab [93] Task A	**Overall approach:** An ensemble of learning algorithms (such as SVM, random forest) fine-tuned using a genetic algorithm. **External resources:** Word features, word embeddings, frequency of emoticons, uppercase characters, among others.
ECNU [104] Tasks A and B	**Overall approach:** A pipeline-based procedure involving relevance and orientation detection. **External resources:** N-grams, topic features and sentiment lexicon features (such as Hu&Liu and MPQA, among others).
CU-GWU [22] Task A	**Overall approach:** Classification using SVM **External resources:** N-grams, Stanford's SA system and LIWC.
IUCL-RF [52] Task A	**Overall approach:** Classification algorithms (SVM, random forest, gradient boosting decision trees) and an ensemble classifier (TiMBL). **External resources:** Bag-of-Words and word vectors.
DeepStance [94] Task A	**Overall approach:** A set of naive bayes classifiers using deep learning. **External resources:** More than 1.5 million of tweets were added by using representative hashtag for target-stance pairs.
UWB [39] Tasks A and B	**Overall approach:** Maximum entropy classifier. **External resources:** N-grams, PoS labels, General Inquirer. Additional tweets were gathered based on frequent hashtags in the training set.
IDI@NTNU [10] Task A	**Overall approach:** A soft voting classifier approach (naive bayes and logistic regression). **External resources:** Word vectors, n-grams, char-grams, negation, punctuation marks, elongated words, among others.
Tohoku [38] Task A	**Overall approach:** Two methods: a feature based approach and a neural network based approach. **External resources:** Bag-of-Words, PoS labels, SentiWordNet. Additional Twitter data was gathered from target words.
ltl.uni-due [100] Tasks A and B	**Overall approach:** Multidimensional classification problem **External resources:** N-grams, punctuation marks, negation, nouns.
JU_NLP [71] Task A	**Overall approach:** Classification using SVM **External resources:** N-Gram and sentiment analysis resources such as: SentiWordNet, EmoLex and NRC Hashtag Emotion Lexicon.
nldsucsc [58] Task A	**Overall approach:** Classification using SVM, J48 and naive bayes. **External resources:** N-grams, PoS labels, LIWC. Additional tweets were gathered based on frequent hashtags in the training set.
INF_UFRGS [19] Task B	**Overall approach:** Set of rules together with SVM. **External resources:** N-grams.
USFD [5] Task B	**Overall approach:** Classification using logistic regression. **External resources:** Bag-of-words autoencoder. Additional tweets were gathered by using two keywords per target.

2.3 Our approach

We are proposing a supervised approach for stance detection [5]. Our work is focused on detecting stance towards Hillary Clinton and Donald Trump that are currently contesting the political campaign for the 2016 U.S. Presidential election. An important aspect to mention concerns to the fact that when the Stance Dataset was built the two targets were still participating to the Party Presidential Primaries for the Democratic and Republican parties, respectively. We address the stance detection in tweets, casting it as a classification task. A set of features that comprises different aspects was exploited. The most novel one refers to the extraction of context-related information regarding to the target of interest. Our hypothesis is that domain knowledge could provide useful information to improve the performance of SD systems. For instance, in order to correctly identify stance in a tweet as the one mentioned in Section 2.2, it is needed to recognize that *Bernie Sander* was an adversary of Hillary Clinton during the Party Presidential Primaries of the Democratic party. Attempting to capture information related to domain knowledge, we define two concepts: “enemies" and “friends". These concepts are used for denoting the entities related to the target. By using the terms “enemies" and “friends", we are trying to infer that when a tweeter is against an “enemy"/“friend" of the target, then the tweeter is in favor/against towards the target and, on the other hand, when a tweeter is in favor towards an “enemy"/“friend" of the target, then the tweeter is against/in favor towards the target. Figure 1 shows an example of the relationships between the “friends" and “enemies" according to their political party, in this case the target of interest is Hillary Clinton.
Three groups of features were considered: sentiment, structural, and context-based.

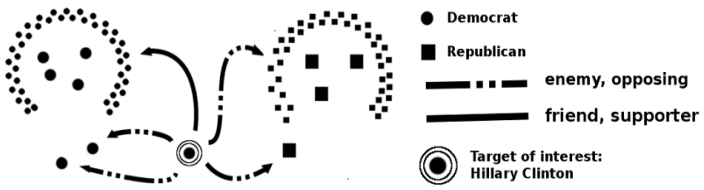

Figure 2.1: Diagram of relationships between friends and enemies of Hillary Clinton

[5] https://github.com/mirkolai/Friends-and-Enemies-of-Clinton-and-Trump

Sentiment-based Features

We shared the idea that stance detection is strongly related to sentiment analysis [60, 104]. As far as we know, there are not sentiment analysis lexica retrieved specifically in the political domain[6]; thus, in order to take advantage of sentiment features it is possible to exploit the wide range of resources available for English. We used a set of four lexica to cover different facets of affect ranging from prior polarity of words to fine-grained emotional information:

- **AFINN**. It is an affective lexicon of 2,477 English words manually labeled with a polarity value between -5 to +5. AFINN was collected by Finn Årup Nielsen [67]. We consider one feature from AFINN: the sum of the polarity of the words present in each tweet.

- **Hu&Liu (HL)**. It includes about 6,800 positive and negative words. We calculate the difference between the positive and negative words in a tweet as a feature.

- **LIWC**. The Linguistic Inquiry and Word Counts (LIWC) [72] is a dictionary that contains about 4,500 entries distributed in 64 categories that can be further used to analyse psycholinguistic features in texts. We calculate the difference between PosEmo (with 405 entries) and NegEmo (with 500 entries) categories in a tweet as a feature.

- **DAL**. The Dictionary of Affect in Language (DAL) contains 8,742 English words; it was developed by Whissell [98]. Each word is rated in a three-point scale into three dimensions: Pleasantness (It refers to the degree of pleasure produced by words), Activation (It refers the degree of response that humans have under an emotional state) and Imagery (It refers to how difficult to form a mental picture of a given word is). We consider six features, i.e. the sum and the mean of the rates of the words present in the tweet for each one of the three dimensions.

Structural Features

We also explore structural characteristics of tweets because we believe that could be useful to detect stance. We experimented with several kinds of structural features, however only the most relevant ones were included in the final approach:

- **Hashtags**. The frequency of hashtags present in each tweet.

[6]For example, the term *vote* is strongly related to politics, but it is not present in commonly used SA lexica such as: AFINN, Hu&Liu, and LIWC.

- **Mentions**. The frequency of screen names (often called mentions) in each tweet.

- **Punctuation marks (punct_marks)**. We consider a set of 6 different features: the frequency of exclamation marks, question marks, periods, commas, semicolons, and finally the sum of all the punctuation marks mentioned before.

Context-based Features

Our hypothesis is that the context-based features should capture some domain-related information. An overall perspective of the context surrounding a target can be acquired by the relationships that exist between the target and other entities in its domain. As mentioned before we are interested in investigating Political debates: for this reason we selected as targets of interest politicians such as Hillary Clinton and Donald Trump. We manually created a list of entities related to the Party Presidential Primaries for the Democratic and Republican parties from Wikipedia[7]. We exploited 6 types of context-based features considering different kinds of relationships between the target and the entities around the target:

- **Target of interest mentioned by name (targetByName)**: This feature captures the presence of the target of interest in the tweet in hand. *#Stop**Hillary**2016 **HillaryClinton** if there was a woman with integrity and honesty I would vote for such as woman president, NO.* The list of tokens used to check the presence of the target of interest are: *hillaryclinton*, *hillary*, *clinton*, and *hill* for Hillary Clinton; while for Donald Trump are *realdonaldtrump*, *donald*, and *trump*.

- **Target of interest mentioned by pronoun (targetByPronoun)**: This feature allows to identify those cases when the target of interest is mentioned by using a pronoun. In the following example, knowing that the target of the tweet is Hillary Clinton, it is possible to exploit the pronoun "she" to capture the presence of the target in hand.

  *HomeOfUncleSam ScotsFyre RWNutjob1 SA_Hartdegen **She**'s too old to understand the internet...that **she** can be fact checked.*
  Two pronouns were considered for each one of the targets of interest: *she* and *her* for Hillary Clinton, while *he* and *his* for Donald Trump.

- **Target's party (targetParty)**: As people involved in politics, our targets belong to a political party. Using this feature we identify if the stance

[7] Articles: *Democratic Party presidential primaries, 2016* and *Republican Party presidential candidates, 2016*

against (or in favor) towards the target of interest was expressed mentioning the name of the party instead of the target. In the following example the tweeter expresses a negative opinion toward Hillary Clinton party.
*It's a miracle, suddenly #**Democrats** don't mind having someone who voted for war.*
In this case we consider the tokens *dem*, *democratic*, *democrat*, *democrats*, *progressive* in order to check the entity party for Hillary Clinton, while we consider the tokens *republican*, *republicans*, and *conservative* for Donald Trump.

- **Party colleague opposite (targetPartyColleagues)**: We also considered the case where the party colleagues of the target of interest are mentioned to express an opinion towards it. We use the name and the surnames of the candidates for the Party Presidential Primaries for both Democratic and Republican parties. In the example, Hillary Clinton's party colleagues are mentioned. *msnbc **Lawrence** JoeBiden Sen**Sanders** we love Joe and **Bernie**–but they ARE too OLD–they would end up a #OneTerm President #SemST*
The list of names used for Hillary Clinton is: *bernie*, *sanders*, *martin*, *o'malley*, *lincoln*, *chafee*, *webb*, *lawrence*, and *lessig*; while for Donald Trump is: *ted*, *cruz*, *marco*, *rubio*, *john*, *kasich*, *ben*, *carson*, *jeb*, *bush*, *rand*, *paul*, *mike*, *huckabee*, *carly*, *fiorina*, *chris*, *christie*, *rick*, *santorum*, *gilmore*, *rick*, *perry*, *scott*, *walker*, *bobby*, *jindal*, *lindsey*, *graham*, *george*, *pataki*.

- **Target's oppositors party (targetsOppositors)**: This feature captures the presence of oppositors belonging to the rival party of target of interest's. In the following example a positive opinion is expressed towards two candidates from the Republican party. Thus, the tweet is against Hillary Clinton.
*PhilGlutting megadreamin Thank you so much for RT and FAV!!! #WakeUpAmerica #**Rubio**2016 #**Cruz**2016 #SemST*
We use the Donald Trump's tokens lists targetParty and targetPartyColleagues in order to create Hillary Clinton's targetsOppositors tokens list, while we use Hillary Clinton's tokens lists targetParty and targetPartyColleagues in order to create Donald Trump's targetsOppositors tokens list.

- **Nobody (nobody)**: This feature allows to catch those cases where any of the above described entities are mentioned in a tweet. In the following example the term *Ambassador* refers to Chris Stevens, who served as the U.S. Ambassador to Libya and who was killed at Bengasi in 2012. The diplomat is related to Hillary Clinton in a situation not related with the

election campaign[8].
I don't want to be appointed to an Ambassador post.
The example also shows how difficult is to infer the stance without a deep knowledge of the context.

After the evaluation of participating systems, the organizers of Semeval-2016 Task 6 annotated the Stance Datataset for sentiment and target in order to explore the relationship between sentiment and stance [59, 60][9]. In particular, tweets were manually annotated by using two additional labels: *Sentiment* and *Opinion Towards*, used to mark the overall sentiment polarity of the tweet and information about the fact that opinion is expressed directly towards the target, respectively:

- **Sentiment**. It can be positive, negative, neutral or none.
- **Opinion_target**. It can take three different values: (1) if a tweet expresses an opinion about the target; (2) if a tweet expresses an opinion related to an aspect of the target or related to something that is not the target; and (3) if there is not opinion expressed.

We decided to exploit such new labels, by enriching our model with corresponding **labeled-based features**, with the aim to experiment with both context and sentiment information provided by human annotators.

2.4 Evaluation

We experimented with a set of tweets belonging to Hillary Clinton and Donald Trump from the Stance Dataset, the Table 2.2 shows the distribution of tweets annotated with stance in the training and the test set for our targets of interest.

Table 2.2: Distribution of stance in training and test set

Targets	% Instances in training				% Instances in test			
	Total	Against	Favor	None	Total	Against	Favor	None
Hillary Clinton	689	57.1	17.1	25.8	295	58.3	15.3	26.4
Donald Trump	-	-	-	-	707	42.3	20.9	36.8

We evaluated our approach by using the same measure defined in [60] in order to compare our results with those participating in the task. We trained a Gaussian Naive Bayes classifier [15] implemented in Scikit-learn Python library[10] to built a model for identifying stance in tweets.

[8] https://en.wikipedia.org/wiki/J._Christopher_Stevens

[9] Notice that this is the first publicly available Twitter dataset annotated with both stance and sentiment.

[10] http://scikit-learn.org/

We adopted two experimental setting: a) **experiment1**. It means to the use of the Sentiment-based, Structural and Context-based features; b) **experiment2**. It refers to the use of all the features described in Section 2.3 including the labeled-based ones. Besides, we experimented using different feature combinations in order to identify which kinds of features could be more relevant for stance detection.

Table 2.3: Best features combination for Hillary Clinton, and the respective results for Donald Trump with experiment1 setting

Feature set	Hillary Clinton			Donald Trump		
	F_{avg}	$F_{against}$	F_{favor}	F_{avg}	$F_{against}$	F_{favor}
mention punct_marks AFINN LIWC HL context_based	63.75	71.95	55.56	53.46	50.29	56.63
punct_marks AFINN LIWC HL context_based	62.70	71.47	53.93	52.76	49.61	55.91
hashtag punct_marks AFINN LIWC HL DAL context_based	62.3	70.43	54.17	50.44	47.69	53.19

Table 2.4: Feature set for Hillary Clinton and the respective results for Donald Trump with experiment2

Feature set	Hillary Clinton			Donald Trump		
	F_{avg}	$F_{against}$	F_{favor}	F_{avg}	$F_{against}$	F_{favor}
hashtag mention context-based labeled-based	71.21	77.17	65.26	69.59	61.99	77.19
hashtag context-based labeled-based	71.02	76.77	65.26	70.40	62.77	78.48
hashtag mention LIWC context-based labeled-based	70.98	78.23	63.73	70.20	63.06	77.35

Tables 2.3 and 2.4 present the best results obtained for Hillary Clinton in the experiment1 and experiment2, respectively. Moreover, those obtained by using the same set of features for Donald Trump are shown. From the results can be noted that the F1-score in “against" class is higher than in “favor". Interestingly, the opposite happens for Donald Trump. The results in Table 2.4 are higher than those from Table 2.3. Table 2.5 shows the best results for Donald Trump using for both experiment1 and experiment2.

Table 2.5: Best feature set for Donald Trump using experiment2 and experiment1 setting

Feature set	Donald Trump		
	F_{avg}	$F_{against}$	F_{favor}
* LIWC HL context_based labeld_based	74.49	69.26	79.72
mention punc_marks HL context_based	55.51	50	61.02

The * indicate the use of features belonging exclusively to experiment2.

As can be noted the context-based features seem to be so relevant for both targets. Besides, it is important to highlight that the best result for

Table 2.6: Results of task A and B

	Task A: Hillary Clinton		Task B: Donald Trump	
	F_{avg}	Ranking	F_{avg}	Ranking
experiment1 for Hillary Clinton	**63.75**	3	53.46	2
experiment1 for Donald Trump	61.25	4	**55.51**	2
experiment2 for Hillary Clinton	**71.21**	1	69.59	1
experiment2 for Donald Trump	68.29	1	**74.49**	1
Systems in the official competition				
INF-UFRGS	-	-	42.32	3
LitisMind	42.08	17	44.66	2
pkudbblab	64.41	2	**56.28**	1
PKULCWM	62.26	3	-	-
TakeLab	**67.12**	1	-	-

each target was not achieved by the same set of features. This is maybe not surprising, if we consider the different political campaign marketing strategies of the two candidates, which can influence also the communication of candidates' oppositors and supporters, both in terms of language register used and addressed topics. For the sake of comparison with the state of the art, we present the results obtained by the three best ranked systems at SemEval-2016 Task 6. We only include the results concerning to Hillary Clinton and Donald Trump. Both the F-measure average and the rank position of each system are included in the Table 2.6. We also show our best results for the two targets using both experimental settings as well as the position in the official ranking in the shared task.

Our approach achieves strongly competitive results. We ranked in the first position for both Task A and Task B using the experiment2 setting considering Hillary Clinton and Donald Trump results. For what concerns to the experiment1 we ranked in the third position for Task A and the second one for Task B. The obtained results outperform the baselines proposed in [60][11]. Besides, our outcomes outrank those obtained by submissions from all teams participating in the shared task (both task A and B). Overall, the results for Hillary Clinton are higher than those for Donald Trump. This was in someway expected, due to the lack of a training set of tweets concerning the target *Donald Trump*.

2.5 Conclusions

In this paper we have shown that including context-related information is crucial in order to improve the performance of stance detection systems. Experiments confirms that stance detection is highly dependent on the domain knowledge of the target in hand. Our approach relies on the presence of entities related to a target in order to try to extract the opinion expressed towards it. Besides, our proposal allows to infer the stance in both cases

[11]The authors experimented with n-grams, char-grams and majority class to establish the baselines for the task.

when the target is explicitly mentioned and also when it is not. The results obtained by exploiting context-related features outperforms those from the best ranked systems in the SemEval-2016 Task 6.

Let us highlight that we are not using either n-grams or any word-based representation, but our approach mainly relies on the context of the target in hand. We plan to investigate the performance of our approach in different domains. Exploiting semantic resources in order to catch additional context information is also an interesting line for future research. Also user's information and her social network structure could be useful. For what concerns to the sentiment-related features, overall results confirm that these kinds of features help in identifying the stance towards a particular target. We exploited different sentiment-related features, ranging from those extracted from affective resources to manually assigned polarity labels.

A further interesting matter of future work could be explore also the stance w.r.t. different aspects of a political target entity. This means to perform a sort of aspect-based sentiment analysis in a political domain, e.g., a tweeter can be in favor of Hillary for aspects related to "Health", but not for other aspects.

Finally, we think that it could be also interesting to investigate how to fruitfully combine information about stance and information about the presence of figurative devices in tweets, such as irony and sarcasm [33, 88], since the use of such devices is very frequent in political debates also in social media and detecting irony and sarcasm have been considered as one of the biggest challenges for sentiment analysis.

Acknowledgments

The National Council for Science and Technology (CONACyT Mexico) has funded the research work of Delia Irazú Hernández Farías (218109/313683). The work of Paolo Rosso has been partially funded by the SomEMBED TIN2015-71147-C2-1-P MINECO research project and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030). The work of Viviana Patti was partially carried out at the Universitat Politècnica de València within the framework of a fellowship of the University of Turin co-funded by Fondazione CRT (World Wide Style Program 2).

Chapter 3

iTACOS at IberEval2017: Detecting Stance in Catalan and Spanish Tweets

Published in:

Lai, M., Cignarella A.T., Hernández Farías D.I. (2017) ITACOS at ibereval2017: Detecting stance in Catalan and Spanish tweets. In: Martínez R., Gonzalo J., Rosso P., Montalvo S., Carrillo-de-Albornoz J. (eds) Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017). Murcia, Spain, September 19, 2017. vol. 1881, pages 185–192. CEUR-WS.
ISSN:1613-0073

Abstract

In this paper we describe the ITACOS submission for the *Stance and Gender Detection in Tweets on Catalan Independence* shared task. Concerning the detection of stance, we ranked as the first position in both languages outperforming the baselines; while in gender detection we ranked as fourth and third for Catalan and Spanish. Our approach is based on three diverse groups of features: stylistic, structural and context-based. We introduced two novel features that exploit significant characteristics conveyed by the presence of Twitter marks and URLs. The results of our experiments are promising and will lead to future tailoring of these two features in a finer grained manner.

3.1 Introduction

Recently, there is a special interest in the task of monitoring people's stance towards particular targets; thus leading to the creation of a novel area of investigation named *Stance Detection* (SD). Research on this topic could have a positive impact on different aspects such as public administration, policy-making, and security. In fact, through the constant monitoring of people's opinion, desires, complaints and beliefs on political agenda or public services, administrators could better meet population's needs. For example, a practical application of SD could improve the automatic identification of people's extremist tendencies (i.e. religious extremism [34]).

In 2016, for the first time a shared task on SD has been held at SemEval-2016, namely the task 6: Detecting Stance in Tweets[1] was organized in the framework of SemEval. The participating teams were required to determine stance towards six different targets: "Atheism", "Climate Change is a Real Concern", "Donald Trump", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". Most of the proposed approaches exploited standard text classification features such as n-grams as well as word embeddings. More details about the participating systems can be found in [8]. In general, related work on SD is scarce, only few works have been published on this novel task. Mohammad et al. [62] took advantage of word-based and sentiment-based features to perform SD on the SemEval-2016 Task 6 dataset. Lai et al. [43], instead, proposed an approach using context features to detect stance towards two targets related to politics in the U.S. presidential elections: Hillary Clinton and Donald Trump. The obtained results outperformed those from the shared task.

In this paper we present our participation to the *Stance and Gender Detection in Tweets on Catalan Independence* task [91] at IberEval-2017[2]. The task is articulated into two subtasks about information contained in Twitter messages written both in Catalan and Spanish: the first subtask is

[1] http://alt.qcri.org/semeval2016/task6/
[2] http://stel.ub.edu/Stance-IberEval2017/

related to detecting author's stance towards the independence of Catalonia, while the second one aims at identifying their gender.

Inferring people's traits such as gender, age or native language on the basis of their written texts is investigated by a field named *Author Profiling* (AP). From 2013 onwards a shared task on AP has been organized at PAN [76, 78, 77, 79] in the framework of CLEF[3]. The intuition behind the task of gender recognition is that of studying how language is used by people and trying to identify features, devices or patterns that are more likely exploited by one gender or the other. More details on the state-of-the-art approaches on this task can be found in [79, 75].

3.2 Our proposal

The starting point of our proposal is to be found in the method proposed in Lai et al. [43] in which the authors exploited three diverse groups of features: *Structural* such as punctuation and other Twitter marks, *Sentiment* i.e. lexica covering different facets of affect, and finally *Context-based*, which consider the relationship that exists between a given target and other entities in its domain.

Therefore, we propose a supervised approach which consists in determining stance towards the independence of Catalonia as well as the gender of the author of a given tweet. In our work, we explored some features that can be grouped in three main categories: *Stylistic*, *Structural*, and *Context*. In the present paper we were not able to explore *Sentiment* features as in [43] due to the fact that we are not aware of sentiment lexica for Spanish and Catalan. We define a set of features distributed as follows:

- **Stylistic Features**
  - Bag of Words (BoW)[4]
  - Bag of Part-of-Speech labels (BoP)[6,5]
  - Bag of Lemmas (BoL)[6,7]
  - Bag of Char-grams (BoC)[6]

- **Structural Features**
  - Bag of Twitter Marks ($BoTM$). We exploit a Bag of Words considering only the words extracted from multi-word Twitter Marks (hashtags and mentions) splitting them by capital letters.

[3] http://clef2017.clef-initiative.eu/

[4] Each tweet was pre-processed for converting it to lowercase. We used unigrams, bigrams and trigrams with a binary representation.

[5] We used TreeTagger [81, 82] for extracting both the part-of-speech and lemmas.

[6] We considered chargrams of 2 and 3 characteres.

- Bag of Hashtags (*BoH*). We consider the hashtags as terms for building a vector with binary representation.
- Frequency of Hashtags (*freqHash*).
- Uppercase Words (*UpW*). This feature refers to the amount of words starting with a capital letter.
- Punctuation Marks (*PM*). We take into account the frequency of dot, comma, semicolon, exclamation and question marks.
- Length (*Length*). Three different features were considered to build a vector: number of words, number of characters, and the average of the length of the words in each tweet.

- **Context Features**
  - Language (*Lan*). We create a vector exploiting the labels ES for Spanish and CA for Catalan provided by the organizer.
  - URL (*Url*). We observed that tweets containing a URL are common in the training dataset. We decided to take advantage of this by considering different aspects extracted from short URLs. First, we identified if the web address of reference is or not reachable. Second, we retrieved the words contained on the web address, then we build a bag-of-words using this information.

3.3 Experiments and Results

The organizers provided a dataset of 8,638 tweets written in Spanish and Catalan labelled with stance (AGAINST, FAVOR, and NEUTRAL) and gender (FEMALE and MALE). For what concerns gender, the distribution is balanced among FEMALE and MALE tweets. Regarding stance, the distribution is skew towards FAVOR for Catalan and skew towards NEUTRAL for Spanish (respectively 30.66% and 29.38%). Similar trends were found in Bosco et al. [11].

It appears, therefore, that language could be a useful feature for stance detection in the Catalan independence debate concerning a region characterized by a strong bilingualism and a smoldering nationalism. In fact, «Language divides and unites us. It [...] impinges upon our identity as individuals, as members of a particular ethnic or national group, and as citizens of a given polity» [57]. We therefore believe that there is a strong correlation between stance and the exploitation of language.

In order to assess the performance of the participating systems, a test set of 2,162 unlabelled tweets was provided, and the two tasks were evaluated separately. Two different evaluation metrics were used: (1) the macro-average of F-score (FAVOR and AGAINST) was used in the case of stance

detection and (2) the accuracy was selected as metric to evaluate the performance in terms of gender identification.

3.3.1 iTACOS experiments

In our experiments, we addressed both stance and gender detection as a classification task. The code is available on github for further exploration and for allow reproducibility of our experiments[7]. We carried out several experiments[8] by combining both the features introduced in Section 3.2 together with a set of classifiers composed by: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LG), Decision Tree (DT), and Multinomial Naïve Bayes (MNB). Besides, we exploited a Majority Voting (MV) strategy considering the different predictions of the above mentioned classifiers as described in Liakata et al. [49]. The features we proposed in section 3.2 were exploited in both the tasks of stance and gender detection, but as it will be better described in the result section, they were specifically tailored for the sole purpose of detecting stance and then they were also applied to gender. For this reason, in the present paper we will focus more on the first subtask, that of stance. We analyzed the obtained results and selected the five combinations of features that showed the best performance for the stance detection task. The resulting sets of features are shown in Table 3.1.

We participated in the shared task with five different runs for each language and each subtask. Table 3.2 shows the obtained results by using both the features and the classifier used in each of the submitted runs.

Table 3.1: Best-ranked sets of features using the training set

Name	Features list
Set_α	BoW, BoL, BoC, Url, BoTM, freqHash, UpW
Set_β	BoW, BoL, BoP, BoC, Url, BoH, freqHash, Length
Set_γ	BoW, BoL, BoP, BoC, Url, freqHash, Lan, Length
Set_δ	BoW, BoL, BoP, BoC, Url, freqHash, PM, Length
Set_ϵ	BoW, BoL, BoP, BoC, Url, BoH, PM, Lan

Table 3.2: Results for stance detection on the training set

Run	**Stance Detection**			**Gender Detection**		
	Features and classifier	F-score		Features and classifier	Accuracy	
		Catalan	*Spanish*		*Catalan*	*Spanish*
iTACOS.1	Set_α + SVM	**0.680**	0.544	Set_ϵ + LR	0.720	0.648
iTACOS.2	Set_ϵ + LR	0.633	0.544	Set_δ + LR	0.722	0.648
iTACOS.3	Set_β + LR	0.625	**0.548**	$\overline{5x5}^{*}$	0.728	0.656
iTACOS.4	$\overline{5x5}^{*}$	0.636	0.530	Set_α + MV	0.719	0.646
iTACOS.5	Set_α + MV	0.657	**0.548**	All Sets** + SVM	0.709	0.636

* The final prediction is the most frequent prediction over the 25 combinations between sets of features and machine learning algorithms.
** The final prediction is the most frequent prediction over the 5 combinations between sets of features and SVM.

[7] https://github.com/mirkolai/iTACOS-at-IberEval2017
[8] A 10-fold cross-validation setting was used.

3.3.2 Official results

We ranked as the first position among 10 participating teams in the subtask of stance detection in both Catalan and Spanish. Table 3.3 shows the official results on the test set. At a first glance, it is possible to observe that our proposed approach seems to perform slightly better in Catalan than in Spanish. Overall, our submissions performed better in Catalan, in fact our five runs ranked among the first 8 positions. In Spanish, on the other hand, our less performing run ranked as the 18^{th} position.

Table 3.3: Official results for stance detection

Catalan			**Spanish**		
Ranking	Run	F-score	Ranking	Run	F-score
1	**iTACOS.2**	**0.4901**	**1**	**iTACOS.1**	**0.4888**
2	iTACOS.1	0.4885	7	iTACOS.2	0.4593
4	iTACOS.3	0.4685	12	iTACOS.3	0.4528
7	iTACOS.4	0.4490	14	iTACOS.4	0.4427
8	iTACOS.5	0.4484	18	iTACOS.5	0.4293

As shown in the table above, the best result in each language was not achieved by the same run. iTACOS.2 performs better for Catalan, while iTACOS.1 for Spanish. The poorer results in both languages were obtained by using iTACOS.4 and iTACOS.5. As expected the best performing runs (iTACOS.1 and iTACOS.2) contain both context-based features, validating the importance of considering contextual information in stance detection tasks. For example, both runs include the feature *Url*. We are interested in evaluating the impact of such feature on the performance. For this reason, we carried out experiments on the training set by applying a modified version of iTACOS.1 and iTACOS.2 removing the *Url* feature. Looking at the results, we observed a drop in the performance of -0.029% for Catalan and of -0.002% for Spanish in iTACOS.1; and of -0,004% for Catalan and of -0.002% for Spanish in iTACOS.2.

The *BoTM*, a novel feature included in the structure-based group, emerges among the relevant features in iTACOS.1 concerning Spanish, but further inquiry on its relevance is matter of future work. For what concerns classifiers, LG and SVM achieved the best performance in both languages. Surprisingly, the approach exploiting MV is not performing.

3.3.3 A linguistic revision

A fundamental part of our approach has been that of manually dealing with data. Being the size of the dataset very large, we were able to visualize only a small portion of tweets. Therefore, we focused on the cases of disagreement between the results obtained with iTACOS.1 and the golden labels provided by the organizers[9]. Below, we report some examples both in Catalan and

[9]The tweets have been extracted from the training set.

Spanish:

1. #elecciones #catalunya #NO #27S https://t.co/oBuTDnUEHj
→ #elecciones #catalunya #NO #27S https://t.co/oBuTDnUEHj

LANGUAGE: CATALAN

GOLDEN LABEL: AGAINST

iTACOS.1: FAVOR

2. Ale @JuntsPelSi, a casa, son solo unas #eleccionescatalanas autonómicas. Mañana a trabajar que es lunes. Seguís teniendo el mismo DNI. #27S
→ @JuntsPelSi, go at home, there is only one autonomous #eleccionescatalanas. Tomorrow, go to work that it'll be Monday. You will have the same DNI (Spanish ID). #27S

LANGUAGE: SPANISH

GOLDEN LABEL: AGAINST

iTACOS.1: FAVOR

3. En estas #eleccionescatalanas de decide una posible independencia y un gobierno que vele por los derechos de su pueblo, VOTA @catsiqueespot
→ In these #eleccionescatalanas we decide for a possible independence and a government that fights for the rights of its population, VOTE @catsiqueespot

LANGUAGE: SPANISH

GOLDEN LABEL: FAVOR

iTACOS.1: AGAINST

Example 1, has been marked as FAVOR from our classifier in (iTACOS.1), probably because of the misleading presence of the token "catalunya", written in Catalan. However, the explicit semantic information carried by the hashtag #NO pointing to AGAINST was ignored, thus leading to a wrong classification. Considering Spanish, example 2 has been appointed as FAVOR instead of AGAINST. The presence of the mention @JuntsPelSi (Catalan independence coalition) could have misdirected our classification. On the other hand, the tweet in example 3 was tagged as AGAINST whereas it should have been FAVOR as we clearly infer from "VOTA @catsiqueespot" and according to the golden labels.

A manual analysis of this kind helped us to shed some light on the relevance of each single feature we exploited and, after having linguistically analyzed them, to choose which features had to be included in our final sets.

3.4 Conclusions

In this paper we presented an overview of the iTACOS submission for the *Stance and Gender Detection in Tweets on Catalan Independence* task at IberEval-2017. We participated by submitting five different runs in the detection of author's stance and gender both in Twitter messages in Catalan

and Spanish. Our approach, chiefly based on *context* and *structural* features, proved to be highly successful concerning the task of stance in both languages, as our system ranked as the first position among ten participating teams. The results show that the addition of two particular features, namely *BoTM* and *Url*, produced a significant contribution to Stance Detection task. In the future, we plan to tailor these two features we used in an even finer grained manner.

Chapter 4

Extracting Graph Topological Information and Users' Opinion

Lai, M., Tambuscio M., Patti V., Ruffo G., Rosso P. (2017) Extracting Graph Topological Information and Users' Opinion. In: Gareth J.F. Jones, Séamus Lawless, Julio Gonzalo, Liadh Kelly, Lorraine Goeuriot, Thomas Mandl, Linda Cappellato, Nicola Ferro (eds) Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017). Lecture Notes in Computer Science. Dublin, Ireland, September 11-14 2017, vol. 10456, pages 112-118. Springer, Cham.
ISBN: 978-3-319-65812-4
DOI: https://doi.org/10.1007/978-3-319-65813-1_10

Abstract

This paper focuses on the role of social relations within social media in the formation of public opinion. We propose to combine the detection of the users' stance towards BREXIT, carried out by content analysis of Twitter messages, and the exploration of their social relations, by relying on social network analysis. The analysis of a novel Twitter corpus on the BREXIT debate, developed for our purposes, shows that like-minded individuals (sharing the same opinion towards the specific issue) are likely belonging to the same social network community. Moreover, opinion driven homophily is exhibited among neighbours. Interestingly, users' stance shows diachronic evolution.

4.1 Introduction

The political public debate is radically changed after the increasing usage of social media in last years. Politicians use them in order to conduct their political campaigns, and to engage users. On the other hand, users interact each other sharing their opinions and beliefs about political agenda or public administration. In this domain, techniques to study and analyse social media users' activity have been gaining importance in recent years, and (now more than ever) automatic approaches are needed in order to deal with this enormous amount of users' generated content. For instance, interest is growing in opinion mining, considered an important task to classify and monitor users' sentiment polarity [70], and in Stance Detection (SD), a finer grained task where the focus is on detecting the orientation *pro* or *con* that users assume within debates towards specific target entity, e.g., a controversial issue [60]. SD could be very useful to probe the citizens' perspective towards particular national and international political issues. Many recent works also suggest the exploitation of users' social community to develop features helping to detect their opinions [101, 17]. To learn more about the role of social relations in the formation of public opinion we address two research questions: first, if individuals that share the same opinion towards a specific issue are likely to belong to the same community [55]; second, if link formation can be better understood in term of homophily (i.e., users with the same opinion are more likely to be connected to each other). We also explore the possibility to have a diachronic evolution in stance, e.g., people changing their stance after some particular events, happening when the debate is still active [26]. Here, we analysed the political discussion in United Kingdom (UK) about the European Union membership referendum, held on June 23rd 2016, commonly known as BREXIT, on Twitter. We showed that our hypotheses are supported by the analysis of real data proposing a new SD annotation scheme that takes into account temporal evolution, and a method for SD based on SVM in order to label the stance of users involved in the discussion.

4.2 Dataset

Data collection. In order to explore social relations and temporal evolution of users' stance, we collected about 5M of English tweets containing the hashtag #brexit using the Twitter Stream API, during the time span between June 22nd and 30th. First, we grouped tweets according to three time intervals, corresponding to relevant clear-cut events related to the referendum, in a short and highly focused time window:

- *"Referendum Day"* - the 24 hours preceding the polling stations closing (between June 22nd at 10:00 p.m. and June 23nd at 10:00 p.m.);
- *"Outcome Day"* - the 24 hours following the formalisation of referendum outcome (between June 24nd at 8:00 a.m. and June 25nd at 8:00 a.m.);
- *"After Pound Falls"* - the 24 hours after the financial markets' turbulence that followed the referendum (between June 28nd at 12:00 p.m. and June 29nd at 12:00 p.m.).

Then, we selected a random sample of 600 users from 5,148 that wrote at least 3 tweets in each time interval. We defined a *triplet* as a collection of three random tweets written by the same user in a given time interval. Finally, we created the TW-BREXIT corpus that consists of 1,800 triplets.

Manual annotation. We employed CrowdFlower[1] to annotate the so-obtained corpus. We asked the human contributors to annotate the user's stance on the target *BREXIT* (i.e. UK exit from EU). In particular, given a triplet posted by an user, they had to infer the user's stance, by choosing between three options:

- *Leave*: if they think that the user is *in favour* of the UK exit from EU;
- *Remain*: if they think that the user supports staying within the EU (i.e. the user is *against* BREXIT);
- *None*: if they could not infer user's stance on BREXIT (e.g., all the messages are unintelligible, or the user do not express any opinion about the target, or the user expresses opinion about the target, but the stance is unclear).

The final TW-BREXIT corpus contains 1,760 labelled triplets in agreement (majority voting)[2].

Social Network. By the *friends/list* Twitter API, we collected the follower list for the 4,548 available[3] users over 5,148 that wrote at least 3

[1] http://www.crowdflower.com

[2] Inter-Annotator Agreement: 65.48. The corpus is available for research purposes.

[3] Some users set privacy in order to hide profile information, while others shut down their profile after the referendum.

tweets in each interval in order to explore users' social network. We obtained a graph where a node represents a user and an edge between two users will exist if one follows the other. The graph consists in 4,114,523 nodes connected by 13,189,524 edges. We then extracted a sub-graph consisting in 198,419 nodes connected by 6,604,298 edges after removing friends having less than 10 relations in order to reduce computational issues.

4.3 Content and Network Analysis

Diachronic evolution of stance. In order to provide insights on temporal evolution, we analysed the label distribution in TW-BREXIT over the three temporal intervals. Not surprisingly, we observe an unbalanced distribution for stance as shown in Table 4.1. We used the hashtag #brexit for collecting data: despite it is apparently a neutral hashtag, a recent study [35] shows that most of tweets containing #brexit were posted by people that expressed stance in favour of Brexit, but since we are not interested in predicting the referendum outcome this bias is not crucial for the next analysis. It is more important to notice that label distribution changes over the time, in particular between "Outcome Day" and "After Pound Falls" phases. Then, we considered the point of view of a single user exploring if her/his own stance changes over time. We found that 57,66% of the users was labelled with the same stance in all the three temporal intervals (37,16% Leave, 15,5% None, 5% Remain). Very interestingly, 42,33% of users' labelled stance changes across different temporal intervals. In particular, 9,5% of users' stance varies from *Leave* (L) to *None* (N) (7% L $\rightarrow$ L $\rightarrow$ N; 2,5% L $\rightarrow$ N $\rightarrow$ N). From these results we cannot infer that users effectively changed opinion, but for sure they express their stance in their tweets in a different way depending on the phase of the political discussion. This is an argument in favour of the hypothesis that stance should be analysed not in isolation but also in a diachronic perspective, which will be matter of future deeper investigations.

Table 4.1: Label distribution over the time

Time span	**Leave**	**Remain**	**None**
Average	961 (51%)	236 (14%)	563 (35%)
Referendum day	55.67%	13.67%	30.67%
Outcome day	55.67%	14%	30.33%
After Pound falls	**50%**	13.67%	**36.33%**

Automatic content analysis: stance detection. We aim to automatically estimate the stance of all users of our dataset in order to explore how the stance is distributed in the social network. Then, we propose a machine learning supervised approach using SVM to annotate the stance s of the remaining 3,948 users, using the following five features computed over a triplet: bag of words (BoW), structural-based (structural), sentiment-based (sentiment) (described in [43]), community-based (community), and

temporal-based (temporal). The community feature returns the community of the user who wrote the triple, while the temporal one, the given time interval of the triplet. The F-Measure $\frac{F_{leave} + F_{remain}}{2}$ obtained by SVM using all the mentioned features is 67% and it overcomes the performance of SVM trained with unigrams (58.25%) and unigrams plus n-grams (60.14%) (baselines proposed by [60]).

Community Detection. Subsequently, we analysed the network topology. Figure 4.1(a) shows the graph plotted by the software Gephi [4] coloured by user's community. The users community's membership was assigned by the Louvain Modularity method [9]. Figure 4.1(b) shows the graph where users have been coloured according the annotated stance computed with SVM. Table 4.2 highlights that the percentage of users' stance in community D is evidently biased towards the stance "Remain"; in communities B, E, and F towards the stance "Leave"; in communities A and C towards the stance "None". The existences of communities so defined in term of stance could allow filter bubble phenomena to occur.

Neighbourhood Overlap. Lastly, we evaluated the stance similarity among couples of connected nodes. Then, we defined users *agreement* as a measure of the likelihood that two users i and j have the same stance (i.e., $s(i) = s(j)$) in the same time interval, and then we explored how the agreement between two users changes depending on the rate of the common neighbours. *Neighbourhood Overlap* (NO) is defined as the number of neighbours that nodes i and j have in common divided by the sum of neighbours of both i and j (not counting i and j themselves):

$$NO(i,j) = \left(\frac{|\{N_i \cap N_j\}|}{|\{N_i \cup N_j\} \setminus \{i,j\}|} \right) \tag{4.1}$$

where N_i and N_j are the sets of neighbours of nodes i and j respectively. Table 4.3 shows how to compute the agreement score $\mathrm{A}_{i,j}$ between i and j. Considering $E_l = \{(i,j) \in E | \wedge NO(i,j) = l\}$ as the subset of edges that are incidents to neighbours of both i and j, when $NO(i,j)$ value is exactly equal to l, we computed the *agreement* A_l related to the NO level l as it follows:

$$\mathrm{A}_l(i,j) = \sum_{i,j|(i,j)\in E_l} \frac{A_{i,j}}{|E_l|}. \tag{4.2}$$

Roughly speaking, through this measure we want to explore if the opinion agreement among users changes accordingly the rate of common neighbourhood. We computed users' agreement in our dataset for each time interval and then we took the averaged value for different values of neighbourhood overlap. Figure 4.2 shows that the agreement between two users increases depending on the percentage of friends. Results showed the tendency of users to associate with similar others according to opinion driven homophily.

[4] http://gephi.org

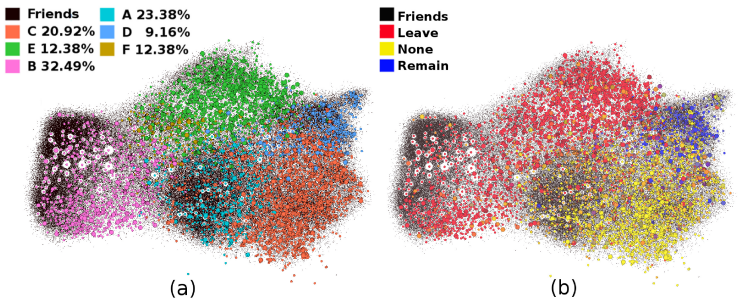

Figure 4.1: In (a), each node is coloured depending on assigned community. Otherwise, in (b), they are coloured according to the annotated stance of the user's triplet by SVM (red for *Leave*, yellow for *None*, blue for *Remain*, mixed colours when stance changes over time). Followers and the remaining users are black-coloured.

Table 4.2: Users' stance distribution over communities. The percentage shows the average users' distribution in communities over the three temporal phases.

Community	A	B	C	D	E	F
Leave	29.63%	**84.61%**	26.31%	18.96%	**85.6%**	**75%**
Remain	11.11%	0.37%	17.02%	**57.47%**	2.37%	0%
None	**56.79%**	14.28%	**54.38%**	18.39%	10.06%	22.92%

Table 4.3: The table shows the Agreement score for couple of users (i, j) over the temporal phases. The maximum value is 1 in the case i and j agree $(s(i) = s(j))$ in all the three temporal phases, 0 if one or both users have label "None" and -1 otherwise.

	Agreement	One or both None	Disagreement
Referendum day	0.33	0	-0.33
Outcome day	0.33	0	-0.33
After Pound falls	0.33	0	-0.33
	1	0	-1

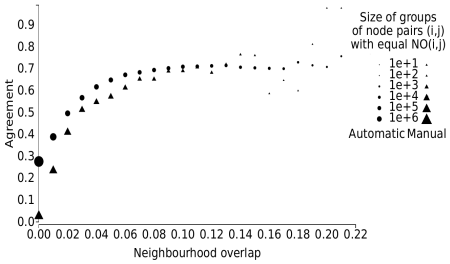

Figure 4.2: A shape (circle or triangle) represents a group of node pairs (i, j) with equal $NO(i, j)$ (rounded to two decimal points). Shape size is proportional to the size of such groups. The agreement score $\mathrm{A}_{i,j}$ was computed with manual annotation stance (triangle) and with user's stance computed by SVM (circle). We noted that the *affinity* among two users increases depending on the rate of NO.

4.4 Discussion

In this paper we have shown that users having the same stance towards a particular issue tend to belong to the same social network community. Moreover, we found evidences that the neighbours are more likely to have similar opinions. The obtained results show that stance verified by human annotators over the same user varies over time, even though we exclusively focused on three 24-hours time slots in a time span of only 8 days. This suggests that stance should be studied considering the diachronic evolution of the debate. We are planning to combine the diachronic evolution of users' stance with the dynamic social network perspective and to explore this methodology on other political corpora. In our future research we would also like to understand the role that influencers could have on the stance change. Moreover, we would like to investigate the use of irony within polarised communities in order to figure out if social network relations influence the use of this figurative language.

Acknowledgments

The work of the last author has been partially funded by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) under the research project SomEMBED TIN2015-71147-C2-1-P and by the Generalitat Valenciana under the grant ALMAMATER (PrometeoII/2014/030).

Chapter 5

Stance Evolution and Twitter Interactions in an Italian Political Debate

Published in:

Lai, M., Patti V., Ruffo G., Rosso P. (2018) Stance Evolution and Twitter Interactions in an Italian Political Debate. In: Silberztein M., Atigui F., Kornyshova E., Métais E., Meziane F. (eds) Proceedings of the 23rd International Conference on Natural Language & Information Systems (NLDB 2018). Lecture Notes in Computer Science. Paris, France, June 13-15 2018, vol. 10859, pages 15-27. Springer, Cham.
ISBN: 978-3-319-91946-1
DOI: https://doi.org/10.1007/978-3-319-91947-8_2

Abstract

The number of communications and messages generated by users on social media platforms has progressively increased in the last years. Therefore, the issue of developing automated systems for a deep analysis of users' generated contents and interactions is becoming increasingly relevant. In particular, when we focus on the domain of online political debates, interest for the automatic classification of users' stance towards a given entity, like a controversial topic or a politician, within a polarized debate is significantly growing. In this paper we propose a new model for stance detection in Twitter, where authors' messages are not considered in isolation, but in a diachronic perspective for shedding light on users' opinion shift dynamics along the temporal axis. Moreover, different types of social network community, based on retweet, quote, and reply relations were analyzed, in order to extract network-based features to be included in our stance detection model. The model has been trained and evaluated on a corpus of Italian tweets where users were discussing on a highly polarized debate in Italy, i.e. the 2016 referendum on the reform of the Italian Constitution. The development of a new annotated corpus for stance is described. Analysis and classification experiments show that network-based features help in detecting stance and confirm the importance of modeling stance in a diachronic perspective.

5.1 Introduction

Nowadays, social media are gaining a very significant role in public debates. Political leaders use social media to directly communicate with the citizens, and citizens often take part in the political discussion, by supporting or criticizing their opinions or proposals. Therefore, social media provide a powerful experimental tool to deduce public opinion's mood and dynamics, to monitor political sentiment, and in particular to detect users' stance towards specific issues, like political elections or reforms, and their evolution during the debate and the related events [13]. Online debates are featured by specific characteristics. As observed by Adamic and Glance, web users tend to belong to social communities segregated along partisan lines [1]. Albeit the scientific debate is still open, some recent studies suggest that the so called "echo chambers" and "filter bubbles" effects tend to reinforce people's pre-existing beliefs, and they also filter and censure divergent ones [92].

In this study we examine the political debate in Twitter about the Italian constitutional referendum held on December 4, 2016 in Italy. To carry on our analysis, we first collected a dataset of about 1M of Italian tweets posted by more than 100K users between November 24 and December 7, 2016, about the Italian constitutional referendum. Then, we extended the collection by retrieving retweets, quotes, and replies, aiming at a representation of political communication through different types of social networks. Furthermore, we divided our dataset in four temporal phases delimited by significant events

occurred around the consultation period, for analyzing the dynamism of both users' stance and social relations. We manually annotated the evolution of the users' stance towards the referendum of 248 users, creating a corpus for stance detection (SD), i.e. the task of automatically determining whether the author of a text is in favour, against, or neutral towards a given target [60]. On this corpus, we were able to analyze the relations that occur among users not only considering the social network structure, but also the users' stance. Based on this analysis we propose a new model for SD in Twitter featured by two main characteristics: (i) network-based features have been included in the model, which result from the analysis of different types of social network communities, based on retweet, quote, and reply relations; (ii) authors' messages are not considered in isolation, but in a diachronic perspective. The major contributions of this work are:

1. *A new resource.* We developed a manually annotated corpus for SD about an Italian political debate, ConRef-STANCE-ita henceforth. Such kind of resource is currently missing for Italian, in spite of the growing interest in the SD witnessed by the recent shared tasks proposed for English [60], Spanish and Catalan [91].

2. *Stance detection.* We propose a new SD model including a set of features based on social network knowledge. Experiments show that analyzing users' relations helps in detecting stance.

3. *Stance diachronic evolution.* Our analysis on the debate provides some evidence that users reveal their stance in different ways depending on the stage of the debate; in particular, our impression is that users tend to be less explicit in expressing their stance as the outcome of the vote approaches.

4. *Network analysis.* Users tend to communicate with similar users, and a strong signal of homophily by stance among supporters and critics of the reform has emerged. Moreover, users having different opinions on the referendum often communicate using *replies*: a significant number of replies posted among ideologically opposed users occurs in the corpus.

The rest of the paper is organized as follows. In Section 5.2 we briefly discuss the related work. In Section 5.3 we describe the development of the corpus, and its characteristics in terms of social network. In Section 5.4 we describe our SD model and the classification experiments. Section 5.5 concludes the paper.

5.2 Related Work

Political sentiment and stance detection. Techniques for sentiment analysis and opinion mining are often exploited to monitor people's mood extracting information from users' generated contents in social media [70]. However, especially when the analysis concerns the political domain [13], a recent trend is to focus on finer-grained tasks, such as SD, where the main aim is detecting users' stance towards a particular target entity. The first shared task on SD in Twitter held at *SemEval 2016*, Task 6 [60], where is described as follows: "Given a tweet text and a target entity (person, organization, movement, policy, etc.), automatic natural language systems must determine whether the tweeter is *in favor* of the target, *against* the given target, or whether neither inference is likely". Standard text classification features such as n-grams and word embedding vectors were exploited by the majority of the participants of the task. The best result was obtained by a deep learning approach based on a recurrent neural network [103].

Machine learning algorithms and deep learning approaches were also exploited in a second shared task held at *IberEval 2017* on gender and SD in tweets on Catalan Independence, with a focus on Spanish and Catalan [91]. With regard to SD, participating teams exploited different kinds of features such as bag of words, bag of parts-of-speech, n-grams, word length, number of words, number of hashtags, number of words starting with capital letters, and so on. The best result was obtained by a support vector machine (SVM) classifier exploiting three groups of features: *Stylistic* (bag of: n-grams, char-grams, part-of-speech labels, and lemmas), *Structural* (hashtags, mentions, uppercase characters, punctuation marks, and the length of the tweet), and *Contextual* (the language of each tweet and information coming from the URL in each tweet) [42].

Political debates and diachronic perspective. Recently, Lai et al. [46] explored stance towards BREXIT at user level by aggregating tweets posted by the same user on 24-hours time windows. This shows how stance may change after relevant events, a finding supported by the work of Messina et al. analysing the same debate [56]. A way to represent a dynamic system is aggregating empirical data over time considering different size of time-windows. Albeit the aggregation time window size is often dictated by the availability of data gathered and this issue has often been neglected in the literature, the importance of the choice of time-windows needs to be considered [40].

Political debate and social media. The huge amount of users generated data allows researchers to observe social phenomena with computational tools in an unprecedented way. Despite social media ease the access to a range of several conflicting views, some works suggest that the existence of the so called "echo chambers" (i.e., when users are exposed only to information from like-minded ones) and "filter bubbles" (i.e., when content is selected

by algorithms according to the user's previous behaviors) can have both positive and negative effects in online and offline forms of political participation [1, 92]. Lazarsfeld and Merto theorized that homophily is involved [47] after the observation that people tend to bond in communities with others who think in similar ways, regardless of any differences in their status characteristics (i.e. gender, age, social status). Recent works shed some light on the relation between social media network structure and sentiment information extracted from posted contents. For example, Lai et. al. [46] reported some preliminary results showing that a strong relation exists between user's stance and friend-based social media community the user belongs, studying the English debate on BREXIT.

5.3 The ConRef-STANCE-ita Corpus

5.3.1 Data Collection and Diachcronic Perspective

Twitter is a microblogging platform where users post short messages called *tweets*. Users can share with their *followers* (users who follow them) the tweets written by other users; this type of shared tweets is known as *retweets*. Furthermore, users can add their own comments before retweeting making a tweet a *quote*. Moreover, it is possible to answer to another person's tweet, generating a so called *reply*. Replying to other replies makes possible the development of longer *conversation threads*, including direct and nested replies.

Researches on Twitter are made easy by the Twitter's REST and Streaming APIs, a set of clearly defined Web services that allow the communication between the Twitter platform and developers. All APIs return a message in JSON, a cross-platform data-interchange format. Also for these reasons, we chose Twitter as platform to gather our experimental data.

Collection.

We collected tweets on topic of the Referendum held in Italy on December 4, 2016, about a reform of the Italian Constitution. On Sunday 4 December 2016, Italians were asked whether they approve a constitutional law that amends the Constitution to reform the composition and powers of the Parliament, the division of powers between the State, the regions, and other administrative entities. This referendum was source of high polarization in Italy and the outcome caused a sort of political earthquake[1]. The data collection consists of four steps:

1. About 900K tweets were collected between Nov. 24th and Dec. 7th

[1]The majority of the voters rejected the reform causing the resignation of Matteo Renzi, the Prime Minister that assumed full responsibility for the referendum defeat.

through the Twitter's Stream API, using as keywords the following hastags: #referendumcostituzionale, #iovotosi, #iovotono[2].

2. The source tweet from each retweet was recovered by exploring the tweet embedded within the JSON field *retweeted_status*. Then, we used the *statuses/retweets/:id* Twitter REST API in order to collect all retweets of the each retweeted tweet present in the dataset.

3. We recovered the quoted tweet of each quote exploring the embedded tweet within the JSON field *quoted_status*.

4. We retrieved *conversation threads* recursively resorting to the Twitter REST API *statuses/show/:id*, by using, as parameter, the *id* specified in the field *in_reply_to_status_id* of each replied tweet.

Through these steps, we have thus enlarged the available number of tweets (more than 2M) w.r.t. those gathered by the Twitter Stream API alone (about 900K). Therefore, we extended the number of possible relations between users (retweets, quotes, and replies) involved in the debate through steps 2, 3 and 4 for deeper analyzing social media networks.

Diachronic Perspective.

Using the same methodology described in [46], we divided the collected tweets in four discrete temporal phases, each one delimited by significant daily spikes of tweets. The spikes correspond to events occurred leading up to the referendum, as it is shown in Figure 5.1. We thus consider the following four 72-hour temporal phases:

- "The Economist" (EC): The newspaper *The Economist* sided with the "yes" campaign of the referendum (tweets retrieved between 2016-11-24 00:00 and 2016-11-26 23:59).

- "Demonstration" (DE): A demonstration supporting the "no" campaign of the referendum had been held in Rome exactly one week before the referendum (tweets retrieved between 2016-11-27 00:00 and 2016-11-29 23:59).

- "TV debates" (TD): The Italian Prime Minister, Matteo Renzi, who supported the "yes" campaign of the referendum, participated to two influential debates on TV (tweets retrieved between 2016-11-30 00:00 and 2016-12-02 23:59).

- "Referendum outcome" (RO): The phase includes the formalization of the referendum outcome, and the resignation of the Italian Prime Minister (tweets between 2016-12-04 00:00 and 2016-12-06 23:59).

[2] #constitutionalreferendum, #Ivoteyes, #Ivoteno

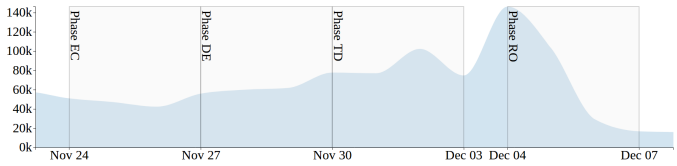

Figure 5.1: Daily frequency of tweets and the discrete division in temporal phases.

5.3.2 Annotation for Stance

We applied to our data the same annotation schema previously exploited at the shared tasks proposed at SemEval 2016 [60] and IberEval 2017 [91] for annotating stance in English, Spanish and Catalan tweets. Here three labels were considered: FAVOR, AGAINST, NONE. The annotation guidelines provided to the annotators follow.

From reading the following tweets, which of the options below is most likely to be true about the tweeter's stance or outlook towards the reform subjected to the Italian Constitutional referendum?

- **FAVOR**: We can infer from the tweet that the tweeter supports the reform.
- **AGAINST**: We can infer from the tweet that the tweeter is against the reform.
- **NONE**: We can infer from the tweet that the tweeter has a neutral stance towards reform or there is no clue in the tweet to reveal the stance of the tweeter towards the reform.

Stance at user level.

We followed the same approach described in [46], where the stance is at user level rather than at tweet level. This means that we deduced the stance from multiple texts written by the same user rather than considering the stance of a single text. We define a *triplet* as a set of three tweets written by the same user in a single temporal phase. The triplet includes: one tweet, one retweet and one reply. This means that each user, for which we annotated the stance, may be a connected node in a network of relations of both retweet or reply. The users who wrote at least one tweet, one retweet, and one reply (a triplet) in each temporal phase are 248. The annotated corpus consists of 992 triplets (248 users by 4 temporal phases). For example, a single user wrote the tweet, the retweet, and the reply highlighted by the black bullet. The reply message in the triplet includes also the related tweet (marked with a white bullet) written by another user.

Tweet	• Travaglio: "Il 2 dicembre grande serata nostra Costituzione in diretta streaming" #ioDicoNo URL via @fattoquotidiano *(Travaglio: "The 2nd December a great night for our Constitution in streaming live" #ISayNo URL through @fattoquotidiano)*
Retweet	•RT @ComitatoDelNO: Brava @GiorgiaMeloni che ricorda a @matteorenzi di (provare a) dire la verità almeno 1 volta su 10! *(RT @NOCommittee: well done @GiorgiaMeloni who reminds to @matteorenzi to (try to) say the truth at least 1 time over 10!)*
Reply	•@angelinascanu @AntonellaGramig @Rainbowit66 per la poltrona. La cosa più cara a voi del #bastaunSi #IoDicoNo #IoVotoNO #vergognaPD *(@angelinascanu @AntonellaGramig @Rainbowit66 for their seats. The most important thing for you of the #justaYES #ISayNo #IVoteNO #shamePD)*
↪ to	◦Già dovrebbe spiegare...ma la risposta si conosce. Il 4 dicembre #bastaunSi #IoVotoSI URL *(He already should explain... but the answer is known. The 4 December #justaYES #IVoteYES URL)*

Manual Annotation.

Two native Italian speakers, domain experts, provided two independent annotations on all the 992 triplets. For what concerns the triplets for which an agreement between the first two annotators was not achieved, we recurred to CrowdFlower[3], a crowd-sourcing platform. We exploited 100 tweets as test questions in order to evaluate the CrowdFlower annotators. We required that annotators were native Italian speakers living in Italy. The annotators have been evaluated over the test questions and only if their precision was above 80% they were included in the task. A further annotator was required unless at least 60% of the previous annotators agreed on the stance of a given triplet. We required a maximum of 3 additional annotators in addition to the 2 domain experts, regarding ambiguous triples. Overall, each triplet was annotated by at least 2 annotators to a maximum of 5.

Agreement.

We calculated the inter-annotation agreement (IAA) as the number of annotators who agree over the majority label divided by the total number of annotators for each single triplet. This type of inter-annotator agreement was proposed by Mohammad et al. [62] to overcome the problem of calculating agreement over a set of documents annotated by a different number of annotators. The IAA calculated over all 992 triplets is 74.6%. Finally, we discarded triplets annotated by 5 annotators having less than 3 annotators in agreement on the same label. We named the Twitter with the stance about the Constitutional reform as ConRef-STANCE-ita, and it consists of 963 triplets.

[3] http://www.crowdflower.com

Label distribution.

Table 5.1 shows the label distribution over temporal phases in the ConRef-STANCE-ita[4]. The percentage of triplets labeled as AGAINST is higher than the rest of labels. This is in tune with the final oucome of the referendum (59.12% vote "no")[5]. The frequency of the label NONE over the different temporal phases is another interesting point. As we can see, the distribution of this label constantly increases from phase EC to phase RO.

Table 5.1: Label distribution

Label	EC	DE	TD	RO	Overall
AGAINST	72.7%	72.7%	71.5%	62.8%	69.9%
FAVOR	19.8%	18.3%	16.9%	14.0%	17.2%
NONE	**6.2%**	**9.1%**	**11.6%**	**22.3%**	12.3%
disagreement	1.2%	0%	0%	0.8%	0.5%

We also explored if users' stance changes over time. We find that 66.8% of the users were labeled with the same stance in all three intervals (55.0% AGAINST, 10.9% FAVOR, 0.8% NONE). For what concerns users that change stance across different time intervals, about 12% of them varies annotated stance in the last phase (10% AGAINST $\rightarrow$ NONE; 2.5% FAVOR $\rightarrow$ NONE). Similar observations were made in [46], while investigating English tweets on the UK European Union membership referendum debate (BREXIT).

5.3.3 Social Media Networks Communities

Networks Science has applications in many disciplines due to networks (or graphs) that are able to represent complex relations among involved actors. Those relations are usually called *edges* and the actors are *nodes*. A network is *weighted* when each edge is characterized with a numerical label that reflects the strength of the connection between two nodes. Therefore, the network is *unweighted* when there is no difference between edges, i.e., all weights are equals to one.

In this work, we represent the relations among Twitter users involved in the Constitutional Referendum debate in the form of graphs. We extracted social media network communities from each graph using the Louvain Modularity algorithm [9]. Then, we examined the structure of four types of communication networks focusing on the dynamism of interactions and the percentage of *uncross-stance* relations (edges between two users with the same stance) for each type of communication. Table 5.2 shows the dimensions of each graph in each temporal phase.

[4]ConRef-STANCE-ita and code available at: github.com/mirkolai/Stance-Evolution-and-Twitter-Interactions.

[5]https://en.wikipedia.org/wiki/Italian_constitutional_referendum,_2016

Table 5.2: Graphs' dimension for each temporal phases.

	RETWEET		QUOTE		REPLY	
	nodes	*edges*	*nodes*	*edges*	*nodes*	*edges*
Overall	94,445	405,843	24,976	69,240	20,936	41,292
EC	25,793	83,134	6,907	13,574	6,236	8,651
DE	28,015	98,717	7,577	15,665	6,663	9,714
TD	33,860	127,593	9,599	22,479	8,801	14,046
RO	63,805	158,243	14,919	21,977	8,497	10,832

Retweet. First, we consider the retweet-based networks. We gathered the retweet list of 649,306 tweets. We created a directed graph for each temporal phase. In particular, an edge between two users exists if one user retweeted a text of the other user during a defined temporal phase. The Louvain Modularity algorithm find about 800 communities for each temporal phase (except for the phase RO where about 1100 communities exist). About 90% of users belong to less than 20 communities.

Quote. We also considered the quote-based networks. We created a directed graph for each temporal phase. An edge between two users exists if one user quotes the other within a defined temporal phase. The four quote-based networks contain about 500 distinct communities (except for phase RO where about 800 communities exist). 1% of the communities contains about 50% of users.

Reply. Finally, we considered the reply-based networks. We recursively gathered the replied tweets of 81,321 replies. The recovered replies are 103,559 at the end of the procedure. Then, we created a directed graph for each temporal phase. In particular, an edge between two users exists if one user replies the other during a defined temporal phase. The communities extracted from the reply-based network are about 700 for each temporal phase (except for phase RO where about 1500 communities exist). There are many communities that contain very few users, indeed only the 2% of the communities contains more than 10 users.

5.3.4 Relations and Stance

Here, we analyze the relations that occur among users not only considering the network structure, but also the users' stance. Table 5.3 shows the percentage of "uncross-stance" relations (edges between two users with the same stance) considering only users annotated with the labels AGAINST or FAVOR. We considered both unweighted and weighted graphs, where the strength of the connection is the number of interactions (retweet, quote, or reply) between two users within the same temporal phase. Following, we evaluated the percentage of "uncross-stance" relations for each of the four network types.

Retweet. First, we analyzed the reply-based network. The considered

Table 5.3: The percentage of uncross-stance relations among users.

	RETWEET		QUOTE		REPLY	
	unweighted	*weighted*	*unweighted*	*weighted*	*unweighted*	*weighted*
Overall	98.6%	99.1%	**94.8%**	**97.6%**	**81.9%**	**77.3%**
EC	98.1%	98.9%	94.0%	96.9%	82.0%	71.9%
DE	99.7%	99.8%	96.1%	97.9%	83.2%	81.0%
TD	98.6%	99.4%	93.9%	97.7%	81.2%	78.9%
RO	**97.5%**	**97.6%**	96.3%	97.9%	80.9%	77.1%

3,099 relations are respectively distributed on the four temporal phases as follows: 749, 885, 989, and 476. The column RETWEET in Table 5.3 shows the percentage of uncross-stance retweets in the retweet-based network. The users usually retweet only tweets belonging to users having the same stance (98.6% and 99.1% overall respectively for unweighted and weighted graphs). There are no significant differences between unweighted and weighted graphs. Notably, the percentage of uncross-stance relations slightly decreases in the phase RO.

Quote. Then, we considered networks based on quote relations. We performed the analysis over 717 relations (respectively 183, 179, 247, and 108 for each temporal phase). The column QUOTE in Table 5.3 shows the percentage of uncross-stance quotes over the temporal phases. There are no significant differences between temporal phases, but the percentage of uncross-stance relations varies between unweighted and weighted graphs (from 94.8% to 97.6% overall).

Reply. Finally, we analyzed the reply-based network. 662 relations are distributed over the four temporal phases as follows: 172, 173, 207, and 110. The column REPLY in Table 5.3 shows the percentage of uncross-stance in both unweighted and weighted for each temporal phase. There are no significant differences between temporal phases, but the percentage of uncross-stance replies significantly varies between unweighted and weighted graphs (in particularly from 81.9% to 77.3% overall). Moreover, here we find a signal that uncross-stance relations is not the whole story.

5.4 Experiments

We propose a new SD model relying on a set of new features, which exploits SVM as machine learning algorithm in a supervised framework. As evaluation metrics, we use two macro-average of the F_{micro} metrics i.e. F_{avg} and $F_{avg_{AF}}$. The first one computes the average among f-AGAINST, f-FAVOR, and F-NONE F_{micro} metrics. The second one, proposed in both SemEval 2016 Task 6 and IberEval 2017 SD tasks [60, 91], computes the average between f-AGAINST and f-FAVOR F_{micro} metrics. We compare our results with two baselines such as: unigrams, bigrams and trigrams Bag of Words using SVM (BoW) and Majority Class ($MClass$). We compute the two met-

rics performing a five-cross validation on the ConRef-STANCE-ita corpus employing each combination of the following features:

- **Bag of Hashtags** (*BoH*) and **Bag of Mentions** (*BoM*): hashtags/mentions as terms for building a vector with binary representation. These features use the texts contained in the tweet, the retweet, and the reply belonging to the triplet.
- **Bag of Hashtags+** (*BoH+*) and **Bag of Mention+** (*BoM+*): tokens extracted from the hashtags/mentions as terms for building a vector with binary representation. We segmented hashtags in tokens using the *greedy algorithm* attempting to find the longest word from a list of about 10M words extracted from Wikipedia's Italian pages. We consider as token the lemma of the verb *to vote* when an inflection of this verb is found. For what concerns mentions, tokens are the result of the *name* splitting, using space as separator. Names are extracted from the *User Object* field *name* of the mentioned user. The feature uses the texts contained in the tweet, the retweet, and the reply belonging to the triplet.
- **Bag of Hashtags+ Replies** (*BoH+R*) and **Bag of Mentions+ Replies** (*BoM+R*): These features are similar to *BoH+* and *BoM+*, but they use information from the conversation thread, by exploiting the text of the replied tweet belonging to the triplet. A different prefix has been used in order to differentiate these tokens from the ones belonging to *BoH+* and *BoM+* feature.

The combination of *BoH+*, *BoM+*, and *BoH+R* (afterwards *TCon*) achieved the highest results, F_{avg} 0.76 and $F_{avg_{AF}}$ 0.85. Notably, removing *BoH+R* from *TCon*, $F_{avg_{AF}}$ declines to 0.83 and F_{avg} declines to 0.69. The model probably is benefiting from the opposition of stance between reply and replied tweets.

Network-based Features.

In order to study the impact of knowledge of the social network for each network's type, we introduced three new features that consider the community which the user belongs to: **Retweet Communities** (*CRet*), **Quote Communities** (*CQuo*), and **Reply Communities** (*CRep*) respectively. In particular, considering the temporal phase $tp \in \{EC, DE, TD, RO\}$, N binary variables exist, one for each of the N detected communities in the retweet-based, quote-based, or reply-based networks. The variable set to one corresponds to the community to which the users who wrote the triplet belongs in the given temporal phase *tp*. Fig. 5.2 shows the combination of the three network-based features with *TCon*. As we can see, the combination of *TCon*, *CRet*, and *CQuo* achieved the highest value for both F_{avg} and

$F_{avg_{AF}}$ (0.79 and 0.90, respectively) by improving the results obtained using only the *TCon* features (0.76 and 0.85, respectively). Nevertheless, adding the *CRep* feature does not improve neither F_{avg} and $F_{avg_{AF}}$.

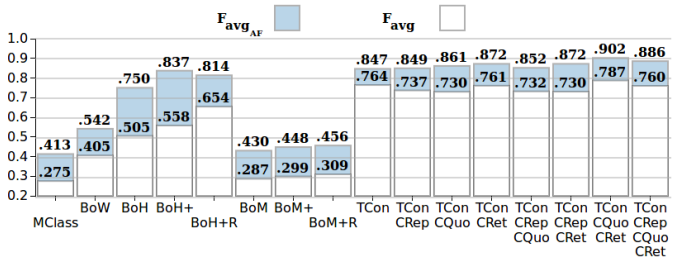

Figure 5.2: F-measured achieved adding network-based features to *TCon*. $F_{avg_{AF}}$: average between f-AGAINST and f-FAVOR F_{micro} metrics. F_{avg}: average among f-AGAINST, f-FAVOR, and F-NONE F_{micro} metrics.

5.5 Discussion and Conclusion

In this work we created a manually annotated Italian corpus for addressing SD from a diachronic perspective, which allows us to shed some light on users' opinion shift dynamics. We observed that in this debate, users tend to be less explicit on their stance as the outcome of the vote approaches. Analyzing the relations among users, we also observed that the retweet-based networks achieved the highest percentage of uncross-stance relations (percentage very close to 100%). This is a signal that Twitter's users retweet almost exclusively tweets they agreed on. Very high percentage of uncross-stance were achieved also by the quote-based networks. The variation between unweighted and weighted graphs could mean that users mainly quote users they agree on. Therefore, it is more likely to be in agreement when the number of quotes connecting two users increases. Interestingly, the opposite is happening on reply-based networks, where we can observe a higher percentage of communications between users with different stances. These observations led us to propose a new model for SD, which includes three new network-based features. The performed experiments show that adding *CRet* and *CQuo* features to content-based features considerably improve the accuracy of SD. We are guessing that when homophily is observed, the user's awareness of being a member of a community can ease user' stance prediction. This does not happen in *CRep*: although the users mainly reply to other users with a similar opinion, we observe about 20% of cross-stance edges. This is a particularly interesting case where inverse homophily (or also heterophily) could be observed. It will be matter of future investigations.

Chapter 6

Discussion of the Results

6.1 Introduction

During my Ph.D., we focused on two main research questions related to polarized political debates. The first one is strictly related to Natural Language Processing and consists in verifying if contextual information (inspecting online social networks) could be useful for detecting the stance towards a target of interest expressed in a piece of text. The second one is a research question related to computational social science which aims to shed some light on dynamics of communications among people having concordant or contrasting opinions, particularly focusing on observing opinions' shifting. We explored several language considering debates on political issues in different countries.

Indeed, our inquiry started from the investigation of a political debate related to the reform process of marriage in France between 2012 and 2013 known as "mariagepourtous" (a law enacted in 2013 for providing the marriage equality between heterosexual and homosexual couples) [41]. In Bosco et al. [12], we describe the creation of a French corpus of tweets annotated for sentiment analysis and for the presence of figurative language devices such as irony and metaphor. The corpus has been developed with the main aim of studying communicative strategies implemented by citizens within political polarized debates on social media. In the meantime, the research community began to highlight the fact that the sentiment extracted from a piece of text is not necessarily equivalent to the opinion expressed towards a particular target of interest (i.e., person, organization, political reform, etc.) [64]. On this basis, a first shared task on stance detection was proposed at SemEval-2016 [60]. We thought stance detection would allow us to better investigate discussions about polarizing topics. For this reason, we decided to focus on the task of stance detection for exploring online polarized debates. A corpus of Spanish and Catalan tweets was developed for sentiment polarity about the independence of Catalonia in Spain [11].

We proposed machine learning approaches for automatically predicting stance and these contributions are discussed in the thesis. In Chapter 2,

we exploited the benchmark dataset of English tweets released at SemEval-2016 [60] focusing on the two main candidates for the 2016 United States presidential primaries for the Democratic and Republican parties: "Hillary Clinton" and "Donald Trump". Here, we proposed a new contextual feature that captures the relations among the involved entities in the debate (DOMAIN KNOWLEDGE) considering the relations of friendship and enmity among the target of interest and the politicians and the parties who have taken part in the electoral campaign. The feature allows us to infer the stance even when the target is not explicitly mentioned, whereas an opinion is expressed towards her friend or enemy. The result obtained by exploiting context-related features is comparable with the current state of the art and outperforms those from the best ranked systems in the SemEval-2016 Task 6.

The second shared task on stance detection held at IberEval-2017 [91] consisted in predicting the stance towards the target "Catalan Indipendence" from Spanish and Catalan tweets. In Chapter 3, we described the iTACOS submission to the *Stance and Gender Detection in Tweets on Catalan Independence* shared task that ranked in the first position in both Spanish and Catalan subtasks. iTACOS employed support vector machine and logistic regression trained with three different groups of features: *stylistic*, *structural* and *context-based*. The results show that two novel features that exploit significant characteristics conveyed by the presence of Twitter marks and URLs produced a significant contribution to stance detection. In particular, the novel feature *URL*, exploring the context of the content of a tweet, identifies the words contained on the shared web address. The feature profits from the fact that users normally tend to share links to web pages which support their position towards the topic. Thus, for example, sharing a web page included in the top-level domain ".cat" (a domain used for promoting the Catalan language and culture) could highlight the "Catalan Indipendence" attitude of the tweeter.

After exploiting features extracted from the textual content of the tweets, we aimed to enrich contextual features exploring the social media network of the users that participated to the debate. Usually, people tend to bond with other having similar opinion, interests, etc. and like-minded people tend to belong to the same social network community [55]. We thought that this observation could be useful for stance detection, thus we proposed to enrich our model with the COMMUNITY feature that takes advantage of knowing the online social community the tweeter belongs to. Benchmark datasets do not allow us to retrieve information about the network of Twitter's users (they usually release only the content of the tweet leaving aside information about the tweeter), for this reason it has been necessary to develop new annotated corpora ourselves. We tested the feature for the first time on *TW-BREXIT*, a dataset for stance detection of English tweets about the United Kingdom's European Union membership referendum (commonly known as

BREXIT) that we created (Chapter 4). We also addressed stance detection in a diachronic perspective proposing an annotation schema that allows to monitor the opinion of the same users through different phases of the debate. In particular, the dataset is divided in three 24-hours steps and each instance corresponds to three tweets (a triplet) written by the same user in the same temporal window. This has allowed us to observe opinions shifting based on the availability of an annotated triplet for each user for each step. We observed that users' labeled stance may consistently change over temporal phases and the label NONE increases approaching to the election date and after the referendum outcome. We speculated that users not only could effectively change opinion, but also they could change their communication style, and probably this affect the annotators' choices. Then, we found evidences that the neighbours are more likely to have similar opinions, and their that have high levels of cross-stance relations, have a higher likelihood to be less explicit (tend to be labeled with label NONE) in expressing her own stance in the next phase of the debate. For extracting features from the users' social networks, we identified the communities from the graph where an edge between two users exists if one follows the other. Then, we carried out classification experiments training a support vector machine with features extracted from the textual content of the triplets (bag of words (BoW), structural-based (*structural*), sentiment-based (*sentiment*)) and using the contextual ones such as community-based (*community*), and temporal-based (*temporal*). From the classification outcomes, it is possible to validate the hypothesis that, using the Louvain Modularity algorithm [9] for extracting the community in which the tweeter belongs to as a feature, allows for improving stance detection prediction.

We also tested our hypothesis on *ConRef-STANCE-ita*, a dateset of Italian tweets for stance detection about the Italian constitutional referendum (Chapter 5). We used the same annotation schema described in Chapter 4, with the only difference that a triplet consists in a tweet, a retweet, and a reply wrote by the same user in the same temporal windows. Then, we widened the size and increased the number of the temporal phases dividing the dataset in four 72-hours steps. The annotation schema allows us to explore different network types and to cover a wider period of time. Indeed, on the one hand, we investigated the friend-based network in the BREXIT debate, on the other hand, the networks based on retweet, quote, and reply have been analyzed here.

Here too, users' labeled stances change over temporal phases, in particular the use of label NONE increases approaching to the referendum outcome. We observed a high percentage of within-stance relations in retweet-based and quoted-based networks and a high percentage of cross-stance relations in networks based on replies. Therefore, it is more likely to be in agreement when users communicate with retweets and quotes. Interestingly, the opposite is happening on reply-based networks, where a higher percentage

of communications between users with different stances exist. We carried out classification experiments training a linear support vector machine with features extracted from the textual content of the triplet (*content-based* features). Then, we extracted three different *community* features, one for each graph i.e., retweet-based, quote-based, and reply-based communities. The performed experiments show that adding features based on retweet and quote relation to *content-based* features considerably improve the accuracy of stance detection prediction. This does not happen using features extracted from the reply-based networks.

In the next sections of this chapter, we present additional experiments we carried out for further investigating what was presented in the previous Chapters. Then, we present results that have not yet been published. First, in Section 6.2, we present an ablation experiment on the features included in iTACOS. Then, in Section 6.3, we present MultiTACOS, a multilingual extension of our stance detection model, that exploits four groups of features such as Stylistic, Structural, Affective and Contextual. In Section 6.4, we performed an ablation test on *TW-BREXIT* corpus for deeper investigating the contribute of the community-based feature on stance detection. Section 6.5 presents an analysis of the users' behaviour on the Italian Constitutional Referendum debate (Chapter 5) shedding some light on the dynamic of communication and polarization among users with diverging point of views. The last section (Section 6.6) proposes to apply network-based features for addressing the task of talent identification. We show as a binary classification task could take advantage of features extracted from a graph representation of the problem and that network-based features can not be exclusively restricted to the task of stance detection.

6.2 An Analysis of the iTACOS Submission at IberEval 2017

We participated to the *Stance and Gender Detection in Tweets on Catalan Independence* shared task held at IberEval 2017 [91]. Our submission achieved the highest result for stance detection for both Catalan and Spanish sub-tasks. In Chapter 3, we described the submitted system iTACOS and the experiments performed with a 10-fold cross validation on the training set. We were unable to carry on feature and error analysis over the test set due to the organizers released it only after the publication of the proceedings of the *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)* [91]. Therefore, we decided to analyze the relevance of each feature used in iTACOS in this section using both training and test sets.

6.2.1 Ablation Experiment in Official Runs

Our submissions consist in five runs performed with the combinations of the following features:

- **Stylistic Features** This group comprises well-known text representations widely used in text classification related tasks. Each tweet was pre-processed for converting it to lowercase. We used TreeTagger [81, 82] for extracting both the part-of-speech and lemmas.

  - Bag of Words (BoW): The feature consists in the binary representation of 1-, 2- and 3- word grams.
  - Bag of Part-of-Speech labels (BoP): The feature consists in the binary representation of 1-, 2- and 3- part of speech grams.
  - Bag of Lemmas (BoL): 1, 2, and 3 lemma grams are considered for implementing this binary feature.
  - Bag of Char-grams (BoC): We considered grams of 2, 3, 4, and 5 chars for extraction a binary representation of the text.

- **Structural Features** We also explore the use of structural characteristics focusing on Twitter Marks.

  - Bag of Twitter Marks ($BoTM$): We exploit a Bag of Words considering only the words extracted from multi-word Twitter Marks (hashtags and mentions) splitting them by capital letters.
  - Bag of Hashtags (BoH): We consider the hashtags as terms for building a vector with binary representation.
  - Frequency of Hashtags ($freqHash$): The feature extracts the number of hashtags contained in the text.
  - Uppercase Words (UpW): This feature refers to the amount of words starting with a capital letter.
  - Punctuation Marks (PM): The feature vector consists in six elements respectively calculating the number of the frequency of dot, comma, semicolon, exclamation, question marks, and the frequency of all punctuation marks.
  - Length ($Length$): Three different elements were considered to build this feature vector: number of words, number of characters, and the average of words length.

- **Context Features** Attempting to take advantage of contextual information, two features were included in this group:

– Language (*Lan*): Due to the nature of the target of interest, the language used by the users could insight their position towards it. We create a binary vector that consists of two elements representing Spanish or Catalan tweets. The language of the tweet is provided by the organizer.

– URL (*Url*): We observed that tweets containing short URLs are common in the training set. We decided to take advantage of this observation by considering different aspects extracted from short URLs. First, we identified whether or not the web address of reference is reachable. Second, we retrieved the web address linked by the short URL. Finally, we built a binary unigram bag-of-words splitting the web address by dot.

We submitted the runs that achieved the highest f-measures performing a 10-fold cross validation on the training set. We take advantage of both linear super vector machine (SVM) and logistic regression (LR) machine learning algorithms. The runs iTACOS.1 and iTACOS.2 achieved the highest official result respectively in Spanish and Catalan subtasks (0.4888). Here, we reported several experiments that remove, from the features configuration of runs iTACOS.1 and iTACOS.2, one feature at time. Table 6.1 shows an ablation test performed in the run iTACOS.1 on Spanish tweets. iTACOS.1 takes advantages of SVM trained with the following features: *BoW*, *BoL*, *BoC*, *Url*, *BoTM*, *freqHash*, *UpW*. We simultaneity reported results obtained by training LR with the same set of features in Table 6.1.

Table 6.1: Ablation test on iTACOS.1

Features	SVM	LR
iTACOS.1	**0.4888**	0.4865
-*BoW*	*0.5197*	0.4826
-*BoL*	0.4591	0.4517
-*BoC*	0.4521	0.4820
-*Url*	*0.4908*	0.4803
-*BoTM*	0.4855	*0.4892*
-*freqHash*	0.4719	0.4813
-*UpW*	0.4859	*0.4891*

Differently from the experiments performed implementing a 10-fold cross validation in the training set, the features *BoW* and *Url* cause a drop in the performance of the system based on SVM when we try to predict the test set. Indeed, iTACOS.1 achieves the highest f-measure removing the feature *BoW* (0.5197) and significantly improves the result removing *Url* (0.4908). The system trained with LR achieves lower results and, using this machine learning algorithm, the performance increases removing *BoTM* or *UpW*.

Table 6.2 shows an ablation test on iTACOS.2. The run iTACOS.2 achieved the highest result in the Catalan sub-task (0.4901) employing LR

trained with the following configuration of features: *BoW*, *BoL*, *BoP*, *BoC*, *Url*, *BoH*, *PM*, *Lan*. Here again, we compared results with those obtained by training SVM with the same set of features.

Table 6.2: Ablation test on iTACOS.2

Features	SVM	LR
iTACOS.2	0.4990	**0.4901**
-*BoW*	0.4712	0.4686
-*BoL*	0.4892	0.4896
-*BoP*	0.4880	*0.5247*
-*BoC*	0.4433	0.4471
-*Url*	0.4804	0.4867
-*BoH*	0.4965	0.4880
-*PM*	0.4966	0.4885
-*Lan*	*0.4990*	0.4876

The highest value is achieved removing the feature *BoP* from iTACOS.2 (0.5247), it means that this feature negatively impacts on the performance of the system. We can highlight the relevance of the feature *BoC* due to the performance significantly decreases removing it (from 0.4901 to 0.4471). Even the Contextual features *Url* and *Lan* positively contribute to the result of iTACOS.2. In this case, the selection of a given machine learning algorithm permits us to achieve highest result (0.4990 using SVM instead of LR). The only feature that does not positively affect results using SVM is the contextual feature *Lan*.

SMV appears to be the best performing machine learning algorithm using the original configuration of runs iTACOS.1 and iTACOS.2. Therefore, we observed that some features negatively impact on the performance and the highest results are obtained discarding *BoW* from iTACOS.1 and *BoP* from iTACOS.2 from the submitted configuration. F-measures overcomes 0.51 in both cases further increasing the distance with the best official results obtained by the participating teams. Although the result overcomes the state of the art, there is always room for improvement due to the fact that F-measure, in absolute terms, is very low.

6.2.2 Evaluating iTACOS Features

The ablation test performed in runs iTACOS.1 and iTACOS.2 shows us that the results obtained with the experiments carried out with a 10-fold cross validation on the training set are significantly different from those obtained predicting the test set. New configurations of features appear to better perform then the ones used in the runs submitted to the shared task. For this reason, we decided to evaluate all features proposed in Chapter 3 including features discarded from the submitted runs. We predict the test set training the machine algorithms with the whole training set. We experimented all combinations of maximum five features and we reported

the highest f-measure obtained for both subtasks. Table 6.3 reports the f-measures obtained adding one feature at time to *BoL* (the configuration that achieved the highest f-measure using only one feature for training SVM) in the Spanish sub-task. The additive test culminates in the configuration of features for both SVM and LR that achieved the best result.

Table 6.3: Additive test in the Spanish sub-task

Spanish			
SVM		*LR*	
BoL	0.4695	*BoL*	0.4917
+BoC	0.4894	*+BoW*	0.4786
++BoTM	0.5010	*++Length*	0.4910
+++freqHash	0.5149	*+++BoH*	0.5105
++++UpW	0.5201		

The highest value is achieved by SVM trained with *BoL*, *BoC*, *BoTM*, *freqHash*, and *UpW* (0.5201). The achieved f-measure continuously increases adding one feature at time to the starting configuration. It is important to note that, using only two features (*BoL* and *BoC*), the corresponding system outperforms the one obtained by the ITACOS.1 run ($0.4894 > 0.4888$).

LR achieved the highest result using only four feature: *BoL*, *BoW*, *length*, and *BoH* (0.5105). Training LR with the only feature *BoL* improves the ITACOS.1 run ($0.4917 > 0.4888$).

Likewise, Table 6.4 reports the F-measures obtained adding one feature at time to *BoW* (the configuration that achieved the highest F-measure using only one feature for training LR) in the Catalan sub-task using both SVM and LR. The final configurations consist of the five features that achieved the best results.

Table 6.4: Additive test in the Catalan sub-task

Catalan			
SVM		LR	
BoW	0.5148	*BoW*	0.4907
+BoC	0.5059	*+BoC*	0.5091
++BoTM	0.5072	*++BoTM*	0.5128
+++Lan	0.5238	*+++Lan*	0.5367
++++Url	0.5531	*++++freqHash*	0.5557

Training SVM with *BoW*, *BoC*, *BoTM*, *lan*, and *url* enables the model to obtain a significant improving of the ITACOS.2 run ($0.5531 > 0.4901$). The contextual features *Lan* and *Url* increase the performances of about 9%. For what concerns LR, the machine learning algorithm trained with *BoW*, *BoC*, *BoTM*, *lan*, and *freqHash* achieves the highest f.measure ($0.5557 > 0.4901$). Here again, the contextual feature *Lan* helps to increase performances of stance detection.

The feature *BoW* is present in all best configurations. Other features based on bag of tokens such as *BoL BoTM*, and *BoC* are also present. Notably, the features based on bag of token are essential for training a machine learning algorithm. Indeed, a team would have easily obtained the best result in the Spanish sub-task if it used the only *BoL* feature for training LG ($0.4917 > 0.4888$). Whereas, to exclusively employ the feature *BoW* for training LR would be enough to obtain the best result in the Catalan sub-task ($0.5148 > 0.4901$).

From previous analysis, the models performing a cross validation on the training set achieve very different results on the test set. For example, we obtained the highest F-measures of 0.680 for Catalan and 0.548 for Spanish performing a 10-cross validation on the training, but the same models achieved lower results on the test set (0.488 for Catalan and 0.453 for Spanish). Potentially, an overfitting occurred and this statistical approach creates a built-in model based on the specific dataset used for training it. Thus, our observation is that, instead of working only to achieve higher F-measures using all kinds of features, we should seek for only proposing features that realistically are consistent with the problem we address. Therefore, the collaboration with a linguist will help to propose more suitable features.

6.3 MultiTACOS: Multilingual Stance Detection

In this section, we propose some machine learning system experiments in a multilingual perspective in order to investigate the portability of our stance detection model across different languages. According to the previous experiences presented in Chapters 2 and 3, we decided to consider datasets of tweets about political polarized debates for training and testing our approach, focusing on targets that are politicians or referendums. Nevertheless, only a few resources annotated for stance currently exist i.e., one in English (described in Chapter 2 and released at SemEval-2016 [60]) and one in Spanish-Catalan (described in Chapter 3 and released at IberEval-2017 [91]). In this section, we respectively call them E-USA and R*-CAT. Moreover, we developed two novel datasets, one for French and one for Italian (E-FRA and R-ITA), for increasing the amount of data available for stance detection. The two novel datasets allow us to enlarge the language scenario making it more adequate for our experiments, but also to make them available to the community research working in this area.

We decided to select topics which are very similar to those of the benchmarks for the purpose of making the novel datasets more comparable with them. The English dataset E-USA focused on the targets related to political elections in the USA ("Hillary Clinton" and "Donald Trump"), thus we collected tweets about the main candidates of the run-off of the French presidential election held in 2017 i.e., "Emmanuel Macron" and "Marine Le Pen"

for the French dataset E-FRA. For what concerns instead the Italian dataset R-ITA, we collected tweets for Italian about the "Constitutional Referendum" held in 2016 in Italy, which mirrors with the target of the Spanish-Catalan corpus "Independence of Catalonia" we redefined as R*-CAT. To avoid any misunderstandings, R-ITA and ConRef-STANCE-ita (Chapter 5) are two distinct corpora. Although they both are composed of tweets gathered during the same political debate (the Italian Constitutional Referendum), they consists of two different samples annotated using two distinct annotation schemes. Indeed, ConRef-STANCE-ita consists of instances annotated for stance composed of a tweet, a retweet and a reply. Whereas a single tweet characterizes each instance of R-ITA.

We following describe the dataset collection, the proposed features and the obtained results.

6.3.1 Data Collection

In this section, we first describe the collection of the SemEval-2016 [60] and IberEval-2017 [91] benchmark datasets and subsequently the collection of the two novel datasets we created for these experiments.

BENCHMARK DATASETS

English Dataset (E-USA) The English dataset is extracted from the prior dataset released by the organizers of the first shared task for stance detection at SemEval-2016 [60]. The organizers gathered tweets using query hashtags concerning the topic of the 2016 United States presidential primaries for the Democratic and Republican parties main candidates, i.e. Hillary Clinton and Donald Trump, such as: *#Hillary4President*, *#Trump2016*, *#WhyIAmNotVotingForHillary*, *#Hillary2016*, *#WakeUpAmerica*. They discarded retweets and tweets with URLs and kept only those where the query hashtags appeared at the end of the tweet. Finally, they removed the query hashtags from each post. From this collection they randomly sampled 2,000 tweets regarding the two candidates that were left after the described preprocessing filtering. See Mohammad et al. [59] for more details about how the dataset was constructed.

Spanish-Catalan Dataset (R*-CAT) The StanceCat dataset was released during the *Stance and Gender Classification Task* that took place as part of IberEval 2017 [91]. Organizers of the shared task used the Stream Twitter API for tracking the hashtags *#Independencia* (#Independence) and *#27S* (27 September, the Catalan regional election day) in order to gather, within September and December 2015, all Spanish and Catalan tweets about the 2015 Catalan regional election that was held on Sunday, 27 September 2015. In total, 10,800 tweets (they discharged retweets) were gathered and

annotated (5,400 written in Catalan and 5,400 written in Spanish). See Taulé et al. [91] for more details about how the dataset was constructed. As a matter of fact, it may be observed that the tweets collected in Spanish-Catalan not properly refer to a referendum. They refer to the "Independence of Catalonia", a subject that has been thoroughly discussed within the 2015 Catalan regional election that was held on Sunday, 27 September 2015, electing the 11th Parliament of the Autonomous Community of Catalonia. An unofficial poll on the same topic, ruled illegal by the Constitutional Court, has been previously held in November 2014, achieving a large majority of votes rooting for independence. According to the view of the secessionists, Catalan regional elections held in September 2015 have been considered a *de facto* referendum on the matter of independence.

NEW DATASETS

French Dataset (E-FRA) We created the French dataset for the present research. It consists of tweets concerning the French presidential elections held in 2017 between the two opponents, i.e. Emmanuel Macron and Marine Le Pen. We used the Twitter Stream API in order to gather about 2,8M tweets (we discarded retweets) over the two weeks preceding and following the second turn of the French presidential elections (held on May 6/7, 2017). The following keywords were used: *macron*, *#presidentielles2017*, *lepen*, and *le pen*. Finally, we randomly selected a sample of 2,000 tweets regarding the figures of Emmanuel Macron and Marine Le Pen.

Italian Dataset (R-ITA) This corpus includes tweets about the topic of the Referendum held in Italy on December 4, 2016, about a reform of the Italian Constitution. On Sunday 4 December 2016, Italians were asked whether they approve a constitutional law that amends the Constitution to reform the composition and powers of the Parliament, the division of powers between the State, the regions, and other administrative entities. We used the Twitter API to gather Italian tweets (excluding retweets) about the debate on this topic tracking the hashtags *#referendumcostituzionale*. We gathered 6K tweets generated by users during the month before the referendum (November 2016). Afterwards, we randomly sampled 1,000 tweets.

The new resources for Italian and French complete the test bed for our experiments about stance detection. The four datasets are indeed featured by comparable topics and size. Nevertheless, the size of the R*-CAT dataset is much bigger than the other three ones. Indeed, an enormous effort has been spent by the organizers of the shared task for building it: it comprise 5,400 annotated tweets for each language.

Data Annotation

All four datasets was annotated following the same guidelines initially proposed for the English dataset in Mohammad et al. [59]. Nevertheless, the intrinsic nature of each language and dataset has determined the application of some minor change in the annotation phase.

In particular, for what concerns the labels of the schema and the criteria to be followed by the annotators for selecting among them in the annotation of each tweet, they are summarized in the following box as reported in [59].

From reading the tweet, which of the options below is most likely to be true about the tweeter's stance or outlook towards the target?

1. **FAVOR**: We can infer from the tweet that the tweeter supports the target.
2. **AGAINST**: We can infer from the tweet that the tweeter is against the target.
3. **NONE**: We can infer from the tweet that the tweeter has a neutral stance towards the target or there is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral) (this label was previously divided in NEUTRAL and NO STANCE) [59].

In the rest of this section we first focus on the two benchmark datasets, and then on the two novel ones, showing the peculiarities of the annotation procedure, guidelines and IAA.

BENCHMARK DATASETS

English Dataset (E-USA) Organizers uploaded the 2,000 tweets of this dataset (1,000 for each of the two targets of stance) on the Crowdflower platform[1] to be annotated by manual annotators. Annotators were previously evaluated against a small gold standard set of previous annotated posts and achieving an accuracy higher than 70%.

The originally proposed annotation schema consists in four labels (i.e. FAVOR, AGAINST, NEUTRAL, and NO STANCE), but they have been reduced to three after the manual annotation took place. Organizers decided to combine NEUTRAL and NO STANCE labels into one unique category, named NONE (neither favor nor against) since less than 0.1% of the data received the NEUTRAL label. After the annotation of each post made by at least eight independent annotators, a corpus including 984 tweets for "Hillary Clinton" and 707 for "Donald Trump" has been released including only tweets having an IAA greater than 60% (at least five out of eight annotators must agree).

The detailed scores of the IAA for the two targets we are interested in ("Donald Trump" and "Hillary Clinton") were not published by the authors. Indeed, the agreement was calculated over all topics and targets (score of 73.11%) as for SemEval-2016 Task 6 comprehended other four targets[2] in addition to "Donald

[1] `https://www.crowdflower.com/`

[2] "Feminist Movement", "Legalization of Abortion", "Atheism' and "Climate Change".

Trump" and "Hillary Clinton". Therefore, the IAA in Mohammad et al. [62] was calculated as the average percentage of times two annotators agreed with each other, with a metric that is not compatible with the most common Fleiss' Kappa coefficient used at IberEval 2017 [91].

Table 6.5: Label distribution in the E-USA dataset

Hillary Clinton				Donald Trump			
FAVOR	*AGAINST*	*NONE*	*TOTAL*	*FAVOR*	*AGAINST*	*NONE*	*TOTAL*
163	**565**	256	984	**299**	148	260	707

Table 6.5 shows the label distribution for each target, in particular, we can see that for the target "Hillary Clinton" a significant unbalanced distribution skewing towards the label AGAINST is present. Whereas the label distribution for "Donald Trump" seems to be skewed towards the label FAVOR.

Spanish - Catalan Dataset (R*-CAT) For building the dataset R*-CAT, released for the IberEval shared task on SD, 5,400 tweets were selected for Catalan and the same amount for Spanish. The annotation schema used for this resource is based on the three labels of the E-USA corpus. The annotation process involved three trained annotators. As first step they tagged stance in 500 tweets in each of the two languages of the corpus and then discussed the annotation in order to achieve agreement and shared guidelines. After that, the three annotators went on to independently annotate the whole corpus. In the released gold resource, one of the labels among AGAINST, FAVOUR or NONE was assigned to a tweet only when proposed by at least two annotators. By contrast, for the tweets on which the three annotators disagreed, the annotation has been discussed until a consensus is achieved at least from two annotators over three. It is important to underline that within this procedure no tweets had been discarded.

Before the consensus was achieved from at least two annotators over the disagreement tweets, the IAA on 10,800 tweets was calculated through Fleiss' Kappa coefficient reaching a value of $\kappa = 0.60$ in both sub-corpora. The results obtained show a moderate agreement, demonstrating the complexity of the task.

Table 6.6: Label distribution in the R*-CAT dataset

Independence of Catalonia (Spanish)				Independence of Catalonia (Catalan)			
FAVOR	*AGAINST*	*NONE*	*TOTAL*	*FAVOR*	*AGAINST*	*NONE*	*TOTAL*
419	1807	**3174**	5,400	**3311**	163	1926	5,400

Table 6.6 shows the label distribution over the two languages for the "Independence of Catalonia" target [3]. As we can appreciate from the numbers shown, a prevalence of the tag NONE features the Spanish posts. On the contrary, tweets written in Catalan have an evident preference for the tag FAVOR. It is also worth

[3] As reported by the organizers of the Stance and Gender Detection in Tweets on Catalan Independence task (at IberEval-2017) [91], the annotators need to know the political context of the debate (e.g. Catalan independence), but is not always sufficient because the text of the tweet leaves room for contrasting interpretations. Organizers tended to assign the neutral NONE labels in these situations. This helps to understand the high number of NONE labeled tweets, particularly high in those written in Spanish.

mentioning the scarce presence of Catalan tweets AGAINST the target "Independence of Catalonia" (only 163 tweets, i.e. 3% of the Catalan sub-corpus). This does not necessarily means that the majority of Catalan people are in FAVOUR of the independence, although the majority of Twitter users writing in Catalan were in FAVOUR.

NEW DATASETS

French Dataset (E-FRA) In the dataset E-FRA we collected tweets in French with the target "Emmanuel Macron" or "Marine Le Pen". The same annotation schema applied for the other datasets has been exploited, but we provided improved guidelines for the label NONE, which has been perceived as especially hard to be annotated. In particular, we detailed the directive for this label as follows: *We can infer from the tweet that the tweeter has a neutral stance towards the target, or there is no clue in the tweet to reveal the stance of the tweeter towards the target (support/against/neutral), or the tweeter considers the target to be the least bad choice*.

The first step of the annotation process consists in the creation of a 100 tweets gold standard for each of the targets by a domain expert. Then native French speakers living in France and achieving an accuracy near to 70% when evaluated against this gold standard. 1,000 tweets for each target are then independently annotated for stance detection by three annotators on CrowdFlower, following the improved guidelines.

The IAA has been separately calculated for each of the two targets. The Fleiss' Kappa coefficient was $\kappa = 0.47$ on tweets targeting "Emmanuel Macron", and $\kappa = 0.44$ on those targeting "Marine Le Pen". Considering this IAA too low, we decided to discard all tweets in which an agreement was not reached by all three. The remaining tweets were 530 for the target "Emmanuel Macron" and 586 for the target "Marine Le Pen".

Table 6.7: Label distribution in the E-FRA dataset

Emmanuel Macron				Marine Le Pen			
FAVOR	*AGAINST*	*NONE*	*TOTAL*	*FAVOR*	*AGAINST*	*NONE*	*TOTAL*
91	**308**	131	530	65	**466**	55	586

Table 6.7 shows the label distribution over the French dataset for both "Emmanuel Macron" and "Marine Le Pen" targets. As we can notice, the label distribution for both targets is skewed towards the label AGAINST.

Italian Dataset (R-ITA) In the dataset R-ITA, the target of interest is the "Constitutional Referendum", and all the tweets are written in Italian. We applied the same annotation process exploited for developing E-FRA, but recruiting native Italian speakers that live in Italy rather than the French ones. The IAA calculated with Fleiss' Kappa coefficient is $\kappa = 0.81$ and demonstrates a substantial agreement (almost perfect) among annotators. The released dataset includes only the 833 tweets obtained by discarding all those not featured by an agreement among all the annotators.

Table 6.8 shows the label distribution over the R-ITA dataset for the target "Constitutional Referendum". As we can notice, the label distribution is skewed towards the label AGAINST.

Table 6.8: Label distribution in the R-ITA dataset

Constitutional Referendum			
FAVOR	*AGAINST*	*NONE*	*TOTAL*
163	**486**	184	833

Table 6.9 shows an overview of the datasets and the distribution of labels for each target. The table contains the number of tweets that overcame all phases of annotation that were not discarded during the process. This is the multilingual test bed we provided for carring out the experiments described in the following sections.

Table 6.9: Overview of label distribution across all datasets

Language	*Target*	*Label distribution*			
		FAVOR	AGAINST	NONE	TOTAL
ENGLISH	Hillary Clinton	163	565	256	984
	Donald Trump	299	148	260	707
SPANISH	Independence of Catalonia	419	1,807	3,174	5,400
CATALAN		3,311	163	1,926	5,400
FRENCH	Emmanuel Macron	91	308	131	530
	Marine Le Pen	65	466	55	586
ITALIAN	Constitutional Referendum	163	486	184	833

6.3.2 Methodology

We used four groups of features such as Stylistic, Structural, Affective and Contextual. We present each group below.

Stylistic Features

First, we pre-processed all the tweets in order to have a lowercase version of them. TreeTagger (`http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/`) [81, 82] was used for extracting both part-of-speech and lemmas. Then, four different text representations were used:

- **Bag of Words** (*BoW*): We considered unigrams, bigrams and trigrams with binary representation.
- **Bag of Part-of-Speech** labels (*BoP*): We considered unigrams, bigrams and trigrams with binary representations of part-of-speech.
- **Bag of Lemmas** (*BoL*): A binary representation of unigrams, bigrams and trigrams of lemmas.
- **Bag of Char-grams** (*BoC*): We exploited a binary representation of chars considering 2, 3, 4, and 5 grams. We included all types of chars, also spaces, dots, commas, etc...

Structural Features

We also explore the use of structural characteristics particularly focusing on Twitter marks.

- **Bag of Twitter Marks** (*BoTM*): We exploited a bag of words considering only hashtags and mentions for creating a binary feature vector.
- **Bag of Hashtags** (*BoH*): We considered the hashtags as terms for building a vector with binary representation.
- **Bag of Hashtags Plus** (*BoHplus*): We considered the terms in the hashtags as tokens for building a vector with binary representation. In this case, we split the hashtag into tokens by capital letters or considering the terms present in the hashtag using a greedy algorithm that takes advantage of a dictionary. We created a dictionary for each language considering the words present in the Wikipedia pages of each election/referendum event[4].
- **Bag of Mention** (*BoM*): We considered the mentions as terms for building a vector with binary representation.
- **Frequency of Hashtags** (*freqHash*): We considered the number of hashtags present in the text.
- **Frequency of Mentions** (*freqMention*): We considered the number of mention tags present in the text.
- **Uppercase Words** (*UpW*): This feature refers to the amount of words starting with a capital letter.
- **Punctuation Marks** (*PM*): We took into account the frequency of dots, commas, semicolons, exclamation, question marks and finally the number of all punctuation marks for creating a feature vector of 6 elements.
- **Length** (*Length*): Three different features were considered: number of words, number of characters, and the average of the length of the words in each tweet.

Affective Features

As it has been investigated in [43, 59], stance detection is strongly related to Sentiment Analysis. Attempting to take advantage of this, we decided to exploit a set of features related to the affective content present in tweets. In doing so, we used different lexical resources defining different kinds of affective information, ranging from overall sentiment to finer-grained aspects (Nissim et al. [68] for a survey on lexical resources for sentiment analysis). Below, we introduce the features we exploited:

- SENTIMENT-RELATED RESOURCES

[4] es.wikipedia.org/wiki/Proceso_participativo_sobre_el_futuro_político_de_Catalu˜na_de_2014,
ca.wikipedia.org/wiki/Consulta_sobre_la_independéncia_de_Catalunya,
it.wikipedia.org/wiki/Referendum_costituzionale_del_2016_in_Italia,
en.wikipedia.org/wiki/United_States_presidential_primary,
en.wikipedia.org/wiki/Democratic_Party_presidential_primaries,_2016,
en.wikipedia.org/wiki/Republican_Party_presidential_primaries,_2016,
fr.wikipedia.org/wiki/élection_présidentielle_française_de_2017

- ***AFINN***: AFFINN [67] is a lexical resource composed by almost 2,500 English words manually annotated with a polarity value in a range from -5 up to +5. It contains a set of words commonly used on the Internet as well as slang acronyms such as LOL (laughing out loud). The feature element contains the sum of the AFINN value of all word present in the text.
- ***HU&LIU***. Hu and Liu [36] proposed two lists of terms related to sentiment (2,006 positive and 4,783 negative words) for opinion mining. This sentiment resource has been widely used by the research community. The feature element contains the sum of the value of all word present in the text (1 if the word is positive, -1 if the word is negative).

- EMOTION-RELATED RESOURCES
  - ***LIWC***. The Linguistic Inquiry and Word Counts (LIWC) is a dictionary developed by Pennebaker et al. [72]. It contains more than four thousands of words distributed in several categories for analyzing psychological aspects in written texts. A category related to emotions is included in this dictionary. The feature element contains the sum of the value of all word present in the text (1 if the word belongs to the category "posemo", -1 if the word belongs to the category "negemo").
  - ***DAL***. Whissel [98] developed the Dictionary of Affect in Language (DAL) which contains 8,742 words annotated on a scale ranging from 0 up to 3 along three dimensions: Pleasantness, Activation, and Imagery. The feature vector contains three elements, one for each dimension. The value of each element is the sum of the values of the dimension of each word present in the text.

All the resources described above have been developed for English. In order to exploit the same set of features in all the other languages we are considering (Spanish, Catalan, Italian and French), we decided to automatically translate the lexical resources via Google Translate APIs. This is a common methodology followed when there are no available resources in languages different from English and in absence of any other language-tailored resource, although sometimes automatic translations are not precise and fully satisfying [2]. The main problem performing an automatically translation of English lexical resources is that it sometimes happens that the original meaning of the word is misinterpreted. For example, the word "agog" in English is positive (+3 in the lexicon AFINN), but Google Translate APIs translate it with the Italian word "impaziente". The meaning of the Italian word "impaziente" is not positive, indeed it correspond to the English words "impatient" or "anxious" that are both evaluated as negative words in the lexicon AFINN (-2).

Contextual Features

Attempting to take advantage of contextual information, three features were included in this group. This kind of information has already proven to be useful in previous stance detection tasks [43]:

- **Language** (*Lan*). Due to the nature of the target of interest, the language could be used as a particular insight on user's position towards it. Here, we

can use this feature only towards the target "Catalan Indipendence" due to the nature of the debate characterized by a request for independence of an autonomous community with a very high percentage of people understanding and speaking both Spanish and Catalan. We created a binary feature vector consisting of two element (Spanish or Catalan) for featuring the language of the tweet.

- **URL** (*Url*). We observed that tweets containing short URLs are common in the datasets. We decided to take advantage of this by considering different pieces of information extracted from the short URL. Firstly, we identified whether the short URL was reachable or not. Second, when available, we retrieved the linked web address. Finally, we built a bag of words using the tokens obtained splitting the web address by dot. Unfortunately, it has not been possible to apply the same procedure to the E-USA dataset, because as explained in Mohammad et al. [59], the tweets containing URLs were discarded in a pre-processing phase.
- **Domain Knowledge** (*Domain*). Lai et al. [43] explored domain knowledge in English tweets concerning Democratic and Republican Parties presidential primaries considering the type of relation among the politicians and parties involved. This feature explored the type of targets' relation between the targets "Hillary Clinton" and "Donald Trump" towards other politicians and parties. We considered the presence of a list of entities in the text and we divided the types of relation and created a binary feature element for each of the following categories:
  - "Target": it identifies the explicit presence of the target (considering the target "Hillary Clinton" the examined keywords were *Hillary* and *Clinton*).
  - "Pronouns": the dataset was created considering only tweets referred to the target, so we considered the presence of a masculine or feminine pronoun referring to the target (considering the target "Hillary Clinton" we looked for the keywords *she* and *her*).
  - "Target's party": the feature identifies the presence of the party that supports the target (for example, the keyword *democratic* for the target "Hillary Clinton" and the keyword *republican* for the target "Donald Trump").
  - "Target's opponent in target's party": the primaries consist in a confrontation between candidates from the same party. The feature identified the presence of at least one member of the target's party (for example, provided that *Bernie Sanders* was candidate against *Hillary Clinton* for the presidency of the democratic party, for this politician we considered the presence of the keywords *bernie* and *sanders*).
  - "Target's opponent in other parties": it considered the candidates for the presidential primaries in the opposite party (for example, provided that *Donald Trump* and *Ted Cruz* were both Republican Party candidates, and that a tweet in FAVOR of a Republican candidate was consequently against the target "Hillary Clinton", that is Democratic, we considered the presence of at least one keyword among *donald*, *trump*, *ted*, and *cruz*).

In this research, we also need to represent and take into consideration the difference of the datasets' domains, i.e. presidential primaries elections and referendums. Therefore, we proposed a modified general set of features verifying the presence of involved entities in the text divided in the following categories:

- "TARGET": the presence of the target (i.e., if the target is "Emmanuel Macron", the presence of the keywords *macron* and *emmanuel* was considered; in the case of "Independence of Catalonia" and "Constitutional Referendum", the keyword *referendum* was considered.
- "TARGET'S SUPPORTERS": the presence of a supporter of the target was considered (e.g. in the case of "Emmanuel Macron" the keyword *brigitte*, Macron's wife; in the case of "Constitutional Referendum" the keywords related to politicians that promoted the reform, like *renzi* or *boschi*, were considered).
- "TARGET'S PARTIES SUPPORTERS": the presence of parties or movements that support the target was considered (i.e., for the target "Catalan indipendence" the presence of keywords referring to the Catalan independence coalition *Junts pel Sí* was considered).
- "TARGET'S OPPONENT": the presence of the target opponents (considering the target "Emmanuel Macron" the keywords related to opposition candidates were considered like e.g. *le pen* and *lepen* were considered).
- "TARGET'S PARTIES OPPONENT": In the last category the presence of the target's opponent party is considered (e.g. provided the target "Constitutional Referendum", the keywords related to the party *Movimento 5 Stelle*, which was against the reform, like *movimento 5 stelle* or *M5S*, were considered).

We made available the full list of keywords for each category and for each target, which was created by a domain expert for each topic at the following link[5].

6.3.3 Experiments

We addressed the stance detection task as a classification problem applying the same strategy previously discussed in Chapters 2 and 3. We exploited for this purpose MultiTACOS for investigating stance detection in a multilingual perspective.

Here, we experimented the use of several supervised learning methods that were already exploited in our previous works obtaining promising results: Linear Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR)[6]. We ran tests with the three methods for each target dataset and in each language using a combination of the 4 groups of features previously described such as *Stylistic*, *Structural*, *Affective*, and *Contextual*[7].

Therefore, we trained 15 models for each proposed machine learning method. For each dataset is provided a 80%-20% split between training and test sets. In

[5]https://github.com/mirkolai/MultilingualStanceDetection

[6]The scikit-learn implementation of the machine learning methods was used (`scikit-learn.org`).

[7]We reported only the results obtained with SVM, and LR because always better than those whit NB.

particular, in the two benchmark datasets (E-USA and R*-CAT) the training set and the test set were released directly from the organizers of the shared tasks, while for the two new datasets (E-FRA and R-ITA) the splitting is randomly performed, maintaining the same ratio of 80%-20% between training and test sets. The macro-average of the F1-score metric (F_{avg} between f-AGAINST and f-FAVOR) proposed at Semeval 2016 [60] and used also at IberEval 2017 [91] was employed to evaluate the prediction of each trained model over the test set.

Considering the benchmark datasets, we compared our results with the results obtained by the best teams competing in each task (SemEval-2016 for E-USA and StanceCat 2017 for R*-CAT [60, 91]). Subsequently we will explore and comment on the experimental results obtained on the new datasets (E-FRA and R-ITA).

Several experiments have been conducted across all datasets, comparing the results in each phase. Our goal was to explore the significance of features in different environments and to test whether the results obtained could be considered language-independent or topic-independent.

BENCHMARK DATASETS

English Dataset (E-USA) We conducted the experiments over the E-USA dataset under a supervised framework for the target "Hillary Clinton" and under a semi-supervised framework for the target "Donald Trump". As we can see from Table 6.10, the best result for "Hillary Clinton" is obtained with a model that exploits SVM as machine learning algorithm trained with *Stylistic*, *Affective*, and *Contextual* features. We trained the model for "Donald Trump" with the tweets about the target "Hillary Clinton" due to the fact that no training set exits for "Donald Trump". The best model for "Donald Trump" exploits LR, but similar results are obtained using SVM. Both the best results with LR and SVM were obtained training the models with *Structural*, *Affective*, and *Contextual* features. As we can notice, both best performing models (results in bold) exploit *Affective* and *Contextual* features.

Table 6.10: The highest F_{avg} values on E-USA dataset

Target	*Classifier*	UNI-GRAM	*Stylistic*	*Structural*	*Affective*	*Contextual*	F_{avg}
Hillary	*LR*	58.18	✓	✓	✓		60.95
Clinton	*SVM*	58.51	✓		✓	✓	**64.51**
Donald	*LR*	21.04		✓	✓	✓	**55.74**
Trump	*SVM*	21.06		✓	✓	✓	55.42

In Table 6.11 we compare the results obtained by our new system with the official results at SemEval-2016 Task 6. As we can see our new system (MultiTACOS) obtains very competitive results (64.51 vs 67.12 and 55.74 vs 56.28). We also include results obtained out of the competition by Lai et al. [43]. We make a distinction between the results obtained by our system on the tweets concerning the target of "Hillary Clinton", for which we scored 64.51 F_{avg} and the results obtained by our system on the tweets concerning the target "Donald Trump" for which the score is 55.74 F_{avg}. The difference of almost 10 points is easily explainable by the

Table 6.11: Our result compared with official results at SemEval-2016 Task 6

Hillary Clinton		
Baselines		
	Majority class	36.83
	SVM-unigrams	57.02
	SVM-ngrams	58.63
	SVM-ngrams-comb	56.50
Participating Teams		
Rank	*Team*	*Result*
1	TAKELAB	67.12
	MULTITACOS	64.51
2	PKUDBLAB	64.41
	Lai et al. [43]	63.65
3	PKULCWM	62.26
4	UWB	59.82
5	IDI@NTNU	57.89

Donald Trump		
Baselines		
	Majority class	29.72
	SVM-ngrams-comb	28.43
Participating Teams		
Rank	*Team*	*Result*
1	PKUDBLAB	56.28
	MULTITACOS	55.74
	Lai et al. [43]	55.51
2	LITISMIND	44.66
3	INF-UFRGS -OPINION-MINING	42.32
4	UWB	42.02
5	ECNU	34.08

fact that the system competing for the automatic stance detection on the target of "Donald Trump" was trained with a training set of tweets concerning another target. For this reason, as we can see, also the performance of all other participating teams is significantly lower. In the same table we also report the scores of the baselines of the shared task: Majority class, SVM-unigrams, SVM-ngrams, and SVM-ngrams-comb[8].

Spanish-Catalan Dataset (R*-CAT) We conducted the experiments over the R*-CAT dataset under the same supervised framework for both languages, training the classifiers on a training set constituted by tweets in both languages.

As we can see from Table 6.12, the best result for the target "Independence of Catalonia" in Spanish is obtained with a model that exploits SVM as machine learning algorithm trained with *Stylistic*, *Structural* and *Affective* features. The best result that our system obtains in Catalan is 48.05 using LR combined with *Structural* and *Affective* features, but it is not enough to reach the results obtained exploiting a system that uses LR trained with the UNI-GRAM baseline which is 50.97. The low results do not come as a surprise, in fact, in StanceCat at IberEval 2017 [91], for the sub-task concerning tweets in Catalan, only one system outperformed

[8]Baselines proposed at SemEval 2016: (1) Majority class: a classifier that simply labels every instance with the majority class ('favor' or 'against') for the corresponding target; (2) SVM-unigrams: five SVM classifiers (one per target) trained on the corresponding training set for the target using word unigram features; (3) SVM-ngrams: a SVM classifier trained using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features; (4) SVM-ngrams-comb: a SVM classifier trained on the combined (all 5 targets proposed in the SemEval-2016 Task 6) training set using word n-grams (1-, 2-, and 3-gram) and character n-grams (2-, 3-, 4-, and 5-gram) features.

the proposed Majority Class LRD baselines[9]. As we can notice, the two best performing models exploit *Affective* and *Structural* features. Additionally the only time that *Contextual* features are used, is for combination with LR in tweets in Spanish.

Table 6.12: The highest F_{avg} values on R*-CAT dataset

Target	*Classifier*	UNI-GRAM	*Stylistic*	*Structural*	*Affective*	*Contextual*	F_{avg}
Catalan Independence (Spanish)	*LR*	44.94	✓	✓	✓	✓	47.78
	SVM	42.01	✓	✓	✓		**48.30**
Catalan Independence (Catalan)	*LR*	**50.97**		✓	✓		48.05
	SVM	46.84	✓	✓	✓		45.89

In Table 6.13, we compare the results obtained by our new system with the official results in StanceCat at IberEval 2017. As we can see our new system obtained top scores both in Spanish and Catalan. The results obtained with the new system (MultiTACOS), developed within the present research, are lower than the ones obtained with the system iTACOS due to the fact that we considered features in an aggregated way in order to have more advantages in a multilingual scenario and better explore the diverse characteristics of the different groups of features. On the other hand the results of the two iTACOS runs are higher because the set of features that we exploited in Lai et al. [42] were specifically tailored for the StanceCat task[10].

NEW DATASETS

French Dataset (E-FRA) We carried out the experiments over the E-FRA dataset under a supervised framework for the target "Emmanuel Macron" and under a semi-supervised framework for the target "Marine Le Pen" with the aim of emulating a procedure similar to the one we used for the E-USA dataset.

As we can see from Table 6.14, the best result for "Emmanuel Macron" is obtained with a model that exploits LR as machine learning algorithm trained with *Stylistic*, and *Structural* features. We can notice that in addition to *Stylistic* and *Structural* features, also *Contextual* features are exploited in the best performing model with SVM.

We trained the model for "Marine Le Pen" with the tweets about the target "Emmanuel Macron". We decided to not create a training set for "Marine Le Pen" as well as no training set exits for "Donald Trump" in the E-USA dataset and we wanted to maintain coherence among datasets of the same typology. The best model for "Marine Le Pen" exploits LR trained with *Affective*, and *Contextual* features.

[9] See Lai et al. [42], and Taulé et al. [91]

[10] In the shared StanceCat task at IberEval 2017 we submitted five runs for stance detection in both languages, i.e. five models for Catalan and five models for Spanish. In Table 6.13, they are listed as iTACOS.1, iTACOS.2, etc... .

Table 6.13: Our result compared with official results at IberEval 2017

Catalan Indipendence (Spanish)			**Catalan Indipendence (Catalan)**		
Baselines			*Baselines*		
	Majority class	44.79		*Majority class*	48.82
	LDR	41.35		*LDR*	43.75
Participating Teams			*Participating Teams*		
Rank	*Team*	*Result*	*Rank*	*Team*	*Result*
1	iTACOS.1	48.88	1	iTACOS.2	49.01
	MultiTACOS	48.05	2	iTACOS.1	48.85
2	LTRC_IIITH .system1	46.79		MultiTACOS	48.30
3	LTRC_IIITH .system4	46.40	3	iTACOS.3	46.85
4	ELIRF-UPV.1	46.37	4	LTRC_IIITH .system1	46.75
5	ELIRF-UPV.2	46.37	5	ARA1337.s1	46.59

Table 6.14: The highest F_{avg} values on E-FRA dataset

Target	*Classifier*	UNI-GRAM	*Stylistic*	*Structural*	*Affective*	*Contextual*	F_{avg}
Emmanuel	*LR*	51.69	✓	✓			**57.24**
Macron	*SVM*	52.57	✓	✓		✓	55.97
Marine	*LR*	38.63			✓	✓	**48.57**
Le Pen	*SVM*	34.52			✓	✓	45.58

In Table 6.14 we can see that the best performing models, with the results shown in bold, exploit different groups of features: *Stylistic* and *Structural* for "Emmanuel Macron" and *Affective* and *Contextual* for "Marine Le Pen". We operate a distinction between the results obtained by our system on the tweets concerning the target of "Emmanuel Macron", for which we scored 57.24 F_{avg} (trained with LR) and the results obtained by our system on the tweets about the target "Marine Le Pen" for which the score is 48.57 F_{avg} (trained with LR). The difference of almost 10 points is not surprising because all the models for the target "Marine Le Pen" were trained with a training set of tweets concerning the other target, "Emmanuel Macron", due to the semi-supervised nature of the task.

Italian Dataset (R-ITA) We conducted the experiments over the R-ITA dataset under a supervised framework. As we can see from Table 6.15, the best result for the target "Constitutional Reform" in Italian is obtained with a model that exploits LR as machine learning method trained with *Stylistic*, and *Structural* features. Surprisingly, *Affective* and *Contextual* features do not appear in neither of the three best results that we report. Our intuition behind this situation lies in the fact that we believe the Italian dataset to be particularly *sui generis* when compared

with the other three. The exploitation of hashtags is wide and coherent in the whole corpus. For instance the hashtags #iovotosì (#Ivoteyes) and #iovotono (#Ivoteno) have been exploited almost in each tweet that we took into consideration, and we believe that just their presence (as boolean value) already is a clear manifestation of stance. For this reason *Stylistic* features such as BAG OF HASHTAGS are already sufficient to reach extremely high F-scores (95.92 F_{avg}), and the exploitation of other *Affective* or *Contextual* features is not needed to obtain higher results.

Table 6.15: The highest F_{avg} values on R-ITA dataset

Target	*Classifier*	UNI-GRAM	*Stylistic*	*Structural*	*Affective*	*Contextual*	F_{avg}
Constitutional	*LR*	94.17	✓	✓			**95.92**
Reform	*SVM*	95.11	✓				95.31

In order to explore the importance of some features and in particular, those who exploit the use of hashtags, we performed a separate experiment removing the polarized hashstag #iovotosì (#Ivoteyes), #iovotono (#Ivoteno), #hovotatosi (#Ivotedyes), #votiamono (#wevoteno) etc. from the text of the R-ITA tweets[11]. After this operation, as showed in Table 6.16, SVM achieved the highest result (83.46 F_{avg}) using *Stylistic*, *Structural*, and *Affective* features. *Affective* features gain a particular significance for SD when explicit information derived from tagging in the tweet goes missing or, in this case, is explicitly removed. It is important to note that also whene completely removing all hashtags[12], SVM trained with *Stylistic*, *Structural*, and *Affective* achieved a high F-measure (78.39 F_{avg}).

Table 6.16: The highest F_{avg} values on R-ITA dataset removing polarized hashtags and all hashtags

Removing	*Classifier*	UNI-GRAM	*Stylistic*	*Structural*	*Affective*	*Contextual*	F_{avg}
Polarized Hashtags	*LR*	72.33	✓	✓	✓		81.73
	SVM	73.04	✓	✓	✓		**83.46**
All Hashtags	*LR*	56.43	✓	✓		✓	74.47
	SVM	61.49	✓	✓	✓		**78.39**

A general conclusion of the analysis of the results is that removing hashtags, obviously decreases the quality of results, but at the same time sheds some light on the importance of *Affective* features in SD, as already showed in Mohammad et al. [62].

6.3.4 Discussion

The experiments we performed allowed us to focus on the behaviour of diverse groups of features in different domains from a multilingual perspective. On the one

[11] We used the following case insensitive regular expression #([a-z]{0,}vot[a-z]{1,}) for removing polarized hashtags.

[12] We used the following regular expressions #(w+) for removing all hashtags.

hand, taking into account the five different languages of our datasets (i.e., English, Spanish, Catalan, French and Italian) we could detect linguistic characteristics peculiar of each language. On the other hand, considering the distinction between political elections and referendums, we could see how the features behave differently in different domains and users' styles for communicating stance towards target entities.

Stylistic features, for instance, obtained good results in all supervised contexts and lower results in semi-supervised contexts. The same applied to *Structural* features which perform better in supervised contexts, especially thanks to features connected with *Twitter marks* (hashtags and mentions) with which, users normally express their stance in a debate.

We also performed an ablation test on each group of features, in order to verify which single feature performs better within the same group over a certain dataset. One result worth mentioning is that, within the *Structural* features, the feature that exploits *Twitter marks* is the one which obtains better results. On the other hand, in the semi-supervised contexts, the best results are obtained using modes which exploit *Affective* and *Contextual* features.

Stylistic and *Structural* features are exploited by almost all best performing models in each one of the five considered languages. We also highlight that *Affective* and *Contextual* features are, in fact, language-independent but in general they produce better results over datasets in which the target is a person (i.e., election datasets: E-USA, E-FRA). Moreover, an ablation test conducted on the *Contextual* group of features demonstrated that the feature COMMON KNOWLEDGE is more relevant in supervised contexts where the target is indeed a person.

Furthermore, in the R-ITA dataset, the URL feature profits from the fact that users normally tend to share links to web pages which support their position towards the referendum. The LANGUAGE feature is particularly discriminating with the target "Independence of Catalonia" where nationalist feelings play a big role and language itself is exploited to convey Catalan independentist attitude.

6.4 The Interplay of Online Social Networks and Users' Stance

In this section, we explore in depth the features exploited in Chapter 4 for automatically estimate users' stance about BREXIT referendum. In particular, we performed an ablation test for observing the best performing features for stance detection using the TW-BREXIT corpus.

6.4.1 Case Study

In Chapter 4, we analyzed the political debate about the European Union membership referendum (BREXIT), held on June 23rd 2016 in United Kingdom. The core idea was to consider the evolution of the user's stance during the debate. To do that, we monitored 600 users in three different time intervals, delimited by relevant events. Furthermore, we demanded that each user wrote at least three tweets (a triplet) in each temporal phase. We explored users' online social communities and we found that the user's stance is strongly related to the social media network community the user belongs to.

We proposed two new contextual features respectively based on social communities (*community-context-based*) and on temporal phases (*diachronic-evolution-context-based*). The *community-context-based* feature returns the community of the user who wrote the triple, while the *diachronic-evolution-context-based* one, the time interval in which the triplet has been posted. The model allows us for automatically estimating the stance towards BREXIT of all users of our dataset in order to explore how the stance is distributed in the online social network. As a result of this analysis, we found evidences that the neighbours are more likely to have similar opinions.

In this section, we aim at performing an ablation test on the model we proposed in Chapter 4 to automatically estimate the stance of the users involved in the debate. In the following subsection, we present the set of features we used for training the machine learning algorithms.

6.4.2 Methodology

Our methodology relies on a novel set of contextual features such as *community-context-based* and *diachronic-evolution-context-based*. In addition, we also exploit the contextual feature *common knowledge*, already introduced in Chapter 2 and in Section 6.3. We finally explore *sentiment-based* and *structural-based* features.

Context-based Features

- *Community features*:

  We observed that users with the same stance tend to belong to the same social media network community. In particular, we assumed that the value homophily is involved [47], considering that Twitter users tend to bond with others who think in a similar way, regardless of any difference in their status characteristics (i.e., gender, age, social status). Intuitively, we created a social network based on friendship relations. In particular, an edge between two Twitter's users exists if one *follows* the other.

  Finally, we extracted the social community each user belongs to using the Louvain Modularity algorithm [9][13]. Figure 6.1 shows the 6 communities extracted by the Louvain Modularity algorithm [14] and the average of the label distribution over the communities of the 600 users' stance resulting from TW-ChronosBrexit.

  We observed that the percentage of users' stance in community D is evidently biased towards the stance REMAIN; in communities B, E, and F towards the stance LEAVE; in communities A and C towards the stance NONE. We also noted that the disagreement among contributors is higher for the community D, maybe because the hashtag #brexit is biased in favor of BREXIT [35] and might have contributed to create more ambiguity during the annotation process in a community mostly composed by users with stance REMAIN. The feature extraction results into the definition of the *community-context-based* feature, a binary feature embedding information on the social community the tweeter belongs to. It consists in a binary feature vector of seven elements

[13] We used the software package NetworkX.

[14] The seventh community includes 195 users that were isolated from the graph after removing nodes with a degree lower than 10.

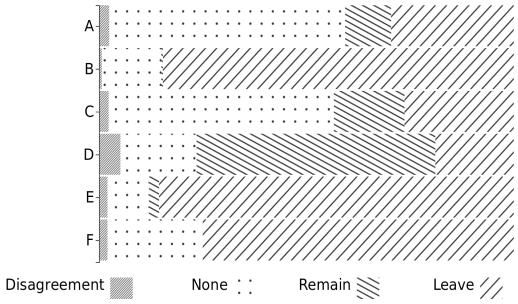

Figure 6.1: The average of the distribution over the communities of the 600 users' stance resulting from the manual annotation process

(one for each community and one for the isolated users) where the element corresponding to the community the user belong to is set to 1.

- *Diachronic evolution features*:

  People's opinion is influenced not only by pre-existent ideology and party identification, but also by information about events happened during the political discussion [26]. Therefore, we hypothize that the evolution of the political debate affects the stance of each voter. Indeed, the dataset was divided in three discrete temporal phases delimited by significant events occurred forthcoming the election day. This assumption does not necessarily imply that users effectively change opinion, but that something changes in the way they write about the topic. The feature *diachronic-evolution-context-based* is so defined as a binary vector taking into account the temporal phase in which the triplet were posted. The binary feature vector consists of three elements (one for each temporal phase) where the element corresponding to the phase the user posts the triplet is set to 1.

- *Common knowledge features*:

  There is a general agreement on the idea that language cannot be investigated in isolation from culture and social organization. Therefore, we decided to consider a feature that takes into account the relation between the target and its supporters or critics. Notably, this feature was already exploited in Chapter 2 and Section 6.3. To address this issue, we first create a gazetteer of parties and politicians involved in the BREXIT debate using both the Wikipedia and DBpedia resources. The gazetteer consists in two values: the name of the party/politician and its/her stance. Secondly, we introduced two binary features considering the mention of politicians and parties entities supporting or not the BREXIT in the text of the tweet: *party-stance-context-based*, a binary feature considering the presence of a mentioned party with its corresponding stance; *politician-stance-context-based*, a binary feature considering the presence of a mentioned politician with her/his corresponding stance. Each of the two feature vectors consists of three elements (party/politician remain, party/politician leave, party/politician neutral) where, for example, the element 'politician remain' is set to 1 if a politician supporting the REMAIN campaign is mentioned in text. Moreover, we introduce another

feature considering the word used for expressing the stance: *explicit-stance-context-based*, a binary bag of word feature vector of length 2 that exclusively considers the words "remain" and "leave".

Sentiment-based Features

Recent works show that stance detection, although having specific characteristics, is strongly related to sentiment analysis [60, 104, 43, 85, 62]. We are not aware of sentiment analysis lexica retrieved specifically in the political domain; thus, we exploited a wide range of resources available for English. We used a set of four lexica to cover different facets of affect, ranging from sentiment polarity of words to fine-grained emotional information: AFINN [67], Hu&Liu [36], LIWC [72], DAL [98]. AFINN was selected since contains several slang and profanity; Hu&Liu and LIWC, since they are widely used in tasks related to analysis of subjective information, and DAL in order to explore different emotional dimensions.

AFINN, Hu&Liu, and LIWC have been exploited for creating three continuous features calculating each value with the average polarity of all words contained in the text. The feature that exploits DAL uses a continuous vector of length three to store the average value of the three dimensions: pleasantness, activation, and imagery.

Structural Features

We also experimented structural characteristics of tweets taking into account the use of metadata and punctuation marks [23]: bag of hashtags, bag of mentions, number of hashtags, number of mentions, and punctuation marks (i.e., frequency of exclamation marks, question marks, periods, commas, semicolons, and, finally, the sum of all the punctuation marks mentioned before).

6.4.3 Experiments

We experimented the use of several supervised learning algorithms such us Naïve Bayes (NB), linear support vector machine (SVM), Random Forest (RF), Decision Trees (DT) on the TW-BREXIT corpus[15]. In addition, we experimented with different feature sets, and evaluated them performing a 5-fold cross validation for each run. We used the macro-average of the F1-score metric and the baselines proposed in Semeval-2016 [60] (such as Majority Class (MC), SVM-unigrams, and SVM-ngrams). The macro-average of the F1-score was redefined, replacing labels FAVOR and AGAINST with labels LEAVE and REMAIN, respectively:

$$F_{\text{avg}} = \frac{F_{LEAVE} + F_{REMAIN}}{2}$$

We experimented stance detection predicting the stance of the user u in the temporal phase t. Here, the classifiers were trained with triplets (three tweets for each user in each temporal phase) and had to predict the stance of the users who wrote the triplet. The union of the text of the three tweets belonging to one triplet is used for calculating the features based on the textual content. We experimented 63 different features combinations using the previously presented six groups of features:

[15] We used the scikit-learn (`http://scikit-learn.org`) implementation of the machine learning algorithms with default parameters.

uni-gram bag of word (*BoW*), structural-based (*structural*), sentiment-based (*sentiment*), common-knowledge-context-based (*comm-know-cxt*), diachronic-evolution-context-based (*de-cxt*), community-context-based (*comm-cxt*). Results are showed in Table 6.17.

Table 6.17: Best feature set on stance at triplet level

Classifier	Feature set	F_{avg}
Baselines		
MC	-	35.25
SVM	unigrams	58.25
SVM	ngrams	60.14
Our Classifiers		
NB	BoW + comm-cxt	53,77
DT	comm-cxt	63.74
RF	comm-cxt	63.76
SVM	structural + sentiment + de-cxt + comm-cxt	**67.01**

The features *Bag of Word* and *structural-based* are relevant due to the presence of three tweets in a triplet (more words than in a text of a single tweet) respectively in Naïve Bayes and SVM. The *community-context-based* feature is significant especially in Decision Tree and Random Forest. In addition, all the best features combinations for each classifier contain the *community-context-based* feature; *de-cxt* feature shows its relevance only in SVM. Table 6.18 shows the results obtained in the ablation test using SVM (i.e., the machine learning algorithm that achieved the best performance in the above mentioned experiments) trained with all the six groups of features. F_{avg} decreases of 14.6% and 0.12% removing singularly *community-context-based* and *diachronic-evolution-context-based* features, respectively. Removing only the *community-context-based* feature F_{avg} improved of 0.49%. Therefore, the *community-context-based* feature does not improve F_{avg} and the *diachronic-evolution-context-based* feature is not decisive in the results. Using the whole group of context based features improves F_{avg} more than using only the *community-context-based* features (16.78%).

Table 6.18: Ablation Test

Features	F_{avg}	Decreasing	Percentage decreasing
All	65.61	0	0%
All - context-based	54.60	-11.01	-16.78%
All - comm-cxt	56.03	-9.58	-14.6%
All - de-cxt	65.53	-0.08	-0.12%
All - comm-know-cxt	65.93	0.32	0.49%
All - sentiment	65.99	0.38	0.58%
All - structural	65.81	0.2	0.3%
All - BoW	65.66	0.05	0.08%

6.4.4 Discussion

Here we investigated the use of several context-based features related to common knowledge, social network community, and diachronic evolution in the stance de-

tection task performing an ablation test on the TW-BREXIT corpus. First, we find that SVM trained using structural, sentiment, diachronic-evolution-context-based, and community-context-based performs better than baselines based on Majority Class, SVM-unigrams, and SVM-ngrams. Results also show that DT and RF achieved high F_{avg} exclusively using the community-context-based feature. Ablation experiments confirm that the entire group of context-based features is very relevant for the stance detection task, in particular the community-context-based one.

However, even if deeper investigations on the possible causes of the opinion shifts are needed, calling also for competencies from other disciplines such as sociology or social psychology, this finding confirms that it is interesting to investigate stance in a diachronic perspective, since opinion fluctuations within the debates occur even in short time spans. It also suggests that people's stance depends not only on their pre-existent ideology and party identification, but also on the information about events happened during the political discussion [26].

6.5 Users' Interactions on Political Debates

In this section, exploring different types of social network communities (i.e., based on friendship, retweet, reply, and quote relations), we aim to analyze the communication among users with similar and divergent viewpoints. We observe particular aspects of users' behaviour in term of peer interaction and opinion shifting, inspecting the case study, particularly designed for exploring users' interactions, described in Chapter 5.

6.5.1 Case Study

Here, we deeply analyze the case study presented in Chapter 5. The dataset consists in more than 2M Italian tweets about the Italian constitutional referendum held in Italy on December 4, 2016. An annotated corpus for stance detection (i.e., ConRef-STANCE-ITA) has been created by monitoring 248 users through the following four temporal phases:

- "The Economist" (EC): The newspaper *The Economist* sided with the "yes" campaign of the referendum (tweets retrieved between 2016-11-24 00:00 and 2016-11-26 23:59).
- "Demonstration" (DE): A demonstration supporting the "no" campaign of the referendum had been held in Rome exactly one week before the referendum (tweets retrieved between 2016-11-27 00:00 and 2016-11-29 23:59).
- "TV debates" (TD): The Italian Prime Minister, Matteo Renzi, who supported the "yes" campaign of the referendum, participated to two influential debates on TV (tweets retrieved between 2016-11-30 00:00 and 2016-12-02 23:59).
- "Referendum outcome" (RO): The phase includes the formalization of the referendum outcome, and the resignation of the Italian Prime Minister (tweets between 2016-12-04 00:00 and 2016-12-06 23:59).

It consists of 992 triplets labeled with one of the following three labels: FAVOR, AGAINST, NONE. We decided to monitor only users that wrote at least one tweet,

one retweet, and one reply (i.e., one triplet) in each temporal phase in a way that at least one friendship, one retweet, and one reply relation exists for each user in each temporal phase.

We also showed that the retweet-based networks achieved the highest percentage of within-stance relations (percentage very close to 100%). In other words, Twitter's users tend to retweet almost exclusively tweets they agreed on. Very high percentage of within-stance were achieved also by the quote-based networks. The variation between unweighted and weighted graphs could mean that users repeatedly quote users they agree on. Interestingly, a higher percentage of cross-stance relations between users with different stances were observed on reply-based networks. Therefore, we propose a new model for stance detection, which includes three new network-based features. The performed experiments show that adding network-based features to content-based ones considerably improves stance prediction in term of both precision and recall.

In this section, we aim at deeply exploring the dynamic of communication among people having concordant and discordant opinion. To do that, we automatically estimate users' stance towards the Italian Constitutional referendum of all users that took part in debates on Twitter. Following, we describe the proposed features used for training the machine learning algorithm.

6.5.2 Methodology

A triplet contains a tweet, a retweet, and a reply (and its replied tweet), therefore we decided to explore features that consider the textual content of each of the three posts. Notably, all proposed features exclusively deal with the textual content due to we do not intentionally exploit network-based features for not creating a dependency between the predicted stance and the structure of the network. Network-based features were been explored and evaluated in Chapter 5. Following, we present the exploited features:

- **Bag of Hashtags** (*BoH*): hashtags considered as terms to build a vector with binary representation.
- **Bag of Mentions** (*BoM*): mentions considered as terms to build a vector with binary representation.
- **Bag of HashtagsPlus** (*BoHplus*): tokens (the longest words found in an Italian dictionary) extracted from the hashtags considered as terms for building a vector with binary representation. The Italian dictionary was created with the most common words extracted from Wikipedia's Italian pages. Particular attention needs to be paid to the verb *to vote*: if the hashtag contains an inflection of this verb we consider the lemma as token.
- **Bag of Mention Plus** (*BoMplus*): tokens extracted from the name of the mentioned users considered as terms for building a vector with binary representation. Names have been extracted from the *User Object* field *name* of the mentioned user, and tokens are the result of the *name* splitting using the space as separator.

Moreover, we also consider other two features that extract a bag of word from the text of the replied tweet (adding a prefix to differentiate the tokens):

- **Bag of Hashtags for Replies** (*BoHplusreply*): same as *BoHplus*, but using the text contained in the replied tweet.
- **Bag of Mentions for Replies** (*BoMplusreply*): same as *BoMplus*, but using the text contained in the replied tweet.

6.5.3 Experiments

We performed a five-cross validation on the ConRef-STANCE-ita corpus with the previous presented features considering the whole triplet.

We used as evaluation metrics, two macro-average of the F_{micro} metrics i.e., F_{avg} and $F_{avg_{AF}}$. The first one computes the average among f-AGAINST, f-FAVOR, and F-NONE F_{micro} metrics. The second one, proposed in both SemEval-2016 Task 6 and IberEval-2017 SD tasks [60, 91], computes the average between f-AGAINST and f-FAVOR F_{micro} metrics. We compared our results with two baselines such as: unigrams, bigrams and trigrams bag of words using SVM (*BoW*) and Majority Class (*MClass*). The combination of *BoHplus*, *BoMplus*, and *BoHplusreply* achieved the highest results ($F_{avg} = 0.76$ and $F_{avg_{AF}} = 0.85$). Both the F_{avg} and $F_{avg_{AF}}$ (see Table 6.19) change in time consistently with IAA (as shown in Table 5.1).

Table 6.19: F_{avg} and $F_{avg_{AF}}$ achieved in the different temporal phases with the combination of BoHplus,BoMplus, and BoHplusreply features.

	Overall	EC	DE	TD	RO
F_{avg}	0.76	0.58	0.72	0.83	0.62
$F_{avg_{AF}}$	0.85	0.87	0.87	0.90	0.72

Table 6.20 shows the F1-score, precision and recall achieved for each class. The model achieved very high values of *Precision* for both AGAINST and FAVOR classes, whereas the class NONE achieved the highest *Recall*.

Table 6.20: Scores achieved by SVM exploiting *BoHplus*, *BoMplus*, and *BoHplusreply*

	NONE	AGAINST	FAVOR
Precision	0.45	**0.96**	**0.94**
Recall	**0.89**	0.86	0.67
F_{micro}	0.60	0.91	0.79

For the sake of completeness, we also report F_{avg} and $F_{avg_{AF}}$ obtained by SVM trained with one of each proposed feature compared with the highest result and baselines as showed in Figure 6.2. We can observe that the feature *BoHplus* achieved an high $F_{avg_{AF}}$, but a relative low $F_{avg_{AF}}$. Furthermore, the feature *BoHplusreply* achieved high values for both $F_{avg_{AF}}$ and $F_{avg_{AF}}$ metrics, but still significantly lower than the highest result.

In order to predict the stance of the increased number of unannotated users that took place in the debate on Twitter, we select all users who wrote at least

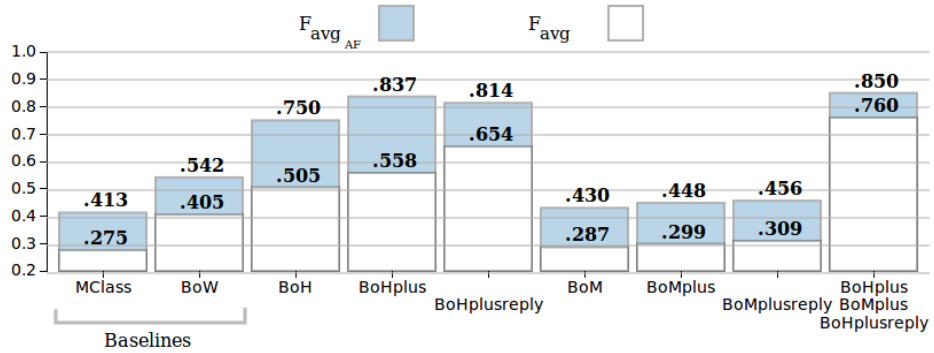

Figure 6.2: F_{avg} and $F_{avg_{AF}}$ obtained by SVM trained with each of the proposed features compared with the baselines and the best feature set result (*BoHplus*, *BoMplus*, and *BoHplusreply*).

one tweet, one retweet and one reply in at least one of the considered temporal phases. We extracted 6441 triplets written by 4,731 different users (excluding users belonging to "Users Sample"). Using the model trained with *BoHplus*, *BoMplus*, and *BoHplusreply*, we automatically annotated the stance of 4,731 different users who were active in at least one temporal phase. Figure 6.3 shows the label distribution in each temporal phase of both manually and automatically annotated triplets.

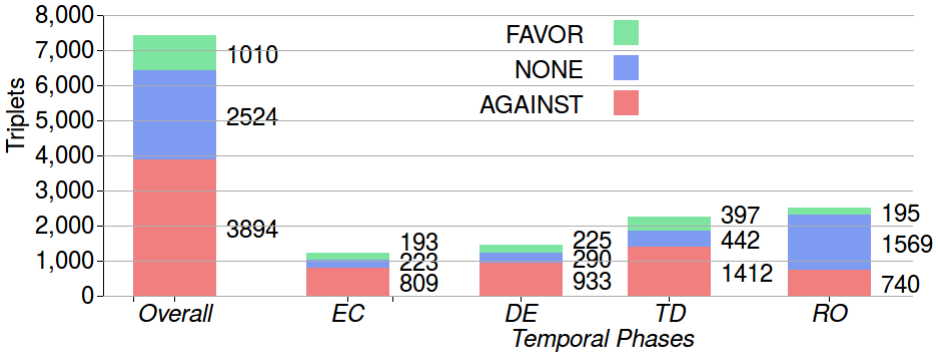

Figure 6.3: Distribution of manually (992) and automatically (6441) annotated triplets over the temporal phases. RO, TD, DE, and EC columns also correspond to the number of labeled users in each phase (one triplet for each user for each temporal phase).

After obtaining the automatically generated annotation for a large number of users, we explored the structure of the four networks respectively based on friends, retweets, quotes and replies. Table 6.21 shows the number of nodes and edges for each network type in each temporal phase.

Table 6.21: Number of nodes and edges for each network type through the four temporal phases.

friend					
	overall	*EC*	*DE*	*TD*	*RO*
nodes	1,383,740	-	-	-	-
edges	5,039,152	-	-	-	-
retweet					
	overall	EC	DE	TD	RO
nodes	94,445	25,793	28,015	33,860	63,805
edges	405,843	83,134	98,717	127,593	158,243
quote					
	overall	*EC*	*DE*	*TD*	*RO*
nodes	24,976	6,907	7,577	9,599	14,919
edges	69,240	13,574	15,665	22,479	21,977
reply					
	overall	*EC*	*DE*	*TD*	*RO*
nodes	20,936	6,236	6,663	8,801	8,497
edges	41,292	8,651	9,714	14,046	10,832

We visualized these networks using the *force atlas* layout[16], hiding users without annotation. The annotated users have been colored depending on the manually or automatically annotated stance: green for FAVOR, red for AGAINST, blue for NONE. For each network, we also included a chord diagram showing the directed inter-relationships among clusters of difference stance.

First, we explore if our graphs exhibit *homophily* according to stance, meaning that users with the same opinion tend to be more connected each other. Let us consider the subnetwork of just FAVOR and AGAINST users. To do this, let A be the fraction of all users annotated as AGAINST and F the fraction of all users annotated as FAVOR. Considering a given edge in any of our four networks, if we assign randomly label AGAINST to the first end of the edge with probability A, and label FAVOR to the other end of the edge with probability F, and vice-versa, than we can have a cross-stance edge with probability $2AF$. Then, applying the *homophily test* proposed in [20], we can just check if the fraction of cross-stance AGAINST-FAVOR edges (CE_{AF}) is significantly less than $2AF$. In such a case, we could conclude that there is a signal of homophily. We can generalize the test including in our observation nodes labelled as NONE. In this case, the probability of a random cross-stance edge is $2(AF + AN + NF)$ (where N is the fraction of all users annotated as NONE). The *homophily test* can be formulated as: "if the fraction of cross-stance edges (CE_{AFN}) is significantly less than $2(AF+AN+NF)$ then there is homophily".

Second, we use *modularity* Q_{AFN} in order to observe the evolution of the polarization among AGAINST, FAVOR and NONE labelled communities during the four temporal phases. Indeed, modularity Q is a network metric that provides a measure of the level of connection among the groups of nodes characterized by

[16]We used the network analysis and visualization framework Gephi (gephi.org)

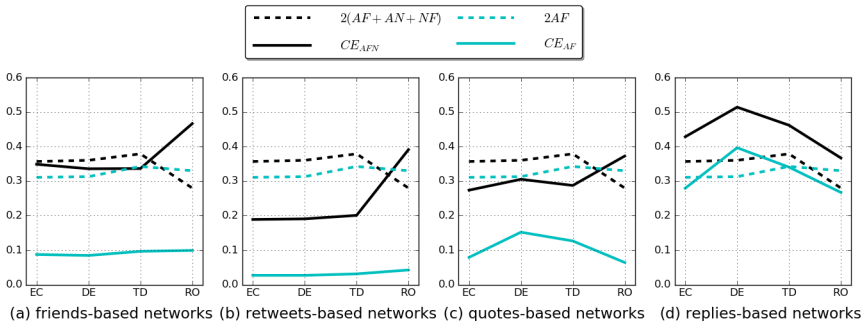

Figure 6.4: The homophily test according to stance for each temporal phase. If the fraction of cross-stance edges observed (solid lines $cross - stance_{AFN}$ and $cross - stance_{AF}$) is significantly less than the probability that a cross-stance link will take place in a null model (dashed lines $2(AF + AN + NF)$ and $2AF$) then there is homophily.

different features, or modules [65]. We compute modularity Q as it follows:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij} \tag{6.1}$$

where m is the total number of edges in the network, A_{ij} the element (i,j) of the adjacency matrix of the network ($A_{ij} = 1$ if there is a link between vertices i and j), k_i and k_j are the degree of nodes i and j. The Kronecker δ_{ij} is 1 if users i and j belong to the same group (i.e., are annotated with the same stance, considering AGAINST, FAVOR and NONE labels) and 0 otherwise (Q_{AFN}). Since we do not really know the opinion of NONE users, we also computed the modularity of the networks considering the subnetworks induced by AGAINST and FAVOR users (Q_{AF}). A value of Q=0 should represent a network with a number of within-community edges not higher than a null model. Values higher than 0 indicate a deviation from randomness. Q approaching to 1 indicates strong community structure (even if values higher then 0.7 are very rare [66]).

Since we observed that users labeled as NONE increased in the last phase, we finally explored the likelihood for users to conceal their opinions (changing their stances from AGAINST or FAVOR to NONE) in function of the level of cross-stance edges in the previous phase. First, we computed the fraction of cross-stance edges ρ_i for each user i in the phase t. Then we measured, for each value of ρ, the fraction of users (with the same value of ρ in the temporal phase t) that change their stance from AGAINST or FAVOR to NONE in the phase $t + 1$.

Friends

We start with the analysis of the friends-based networks. Figure 6.7 shows the evolution of the friends-based networks along the four temporal phases. Please notice that the graph structure is the same due to we retrieved the friends list only once after the referendum, but the stance of the single users may change.

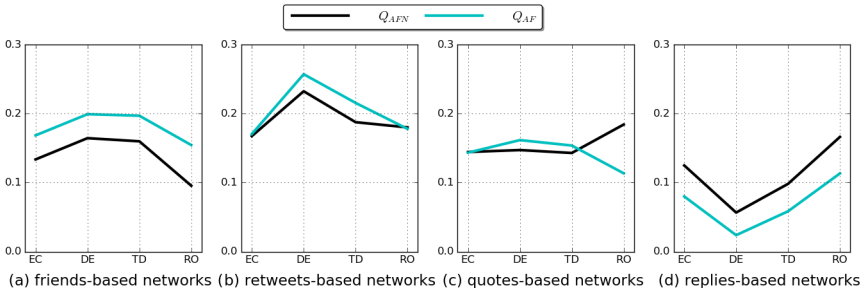

Figure 6.5: Evolution of modularity considering the whole networks (Q_{AFN}) and considering the subnetwork formed by AGAINST and FAVOR clusters (Q_{AF}) for each temporal phase.

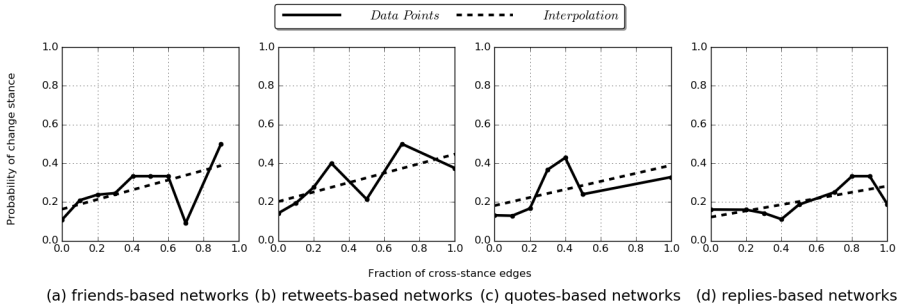

Figure 6.6: The likelihood to change from AGAINST or FAVOR to NONE in function of the fraction of cross-stance edges in the previous phase, for each type of network.

We see segregated colored clusters within the network. The number of users annotated with the label NONE significantly increases in time, in particular in the phase RO. The stance variation seems to affect both AGAINST and FAVOR communities. This is even clearer in the chord diagram of RO phase: the NONE users increase the connections with other groups, but also among themselves.

The graphs do not exhibit homophily by stance (considering the mean and the standard deviation over the four temporal phases, we have a fraction of CE_{AFN} equal to $0.372 \pm \sigma 0.055$ that is slightly higher then $2(AF+AN+NF)$ $0.344 \pm \sigma 0.038$). This means that we have almost a number of cross stance edge that we could expect in a random network with the same characteristics.

However, if we consider each phase separately, we can observe a diverging trend in phase RO. Indeed a quite strong *inverse* homophily by stance emerges among the three clusters (the fraction of CE_{AFN} of 0.467 is significantly higher than $2(AF + AN + NF)$ 0.280) as showed in Figure 6.4(a). As we observe in the chord diagram, this means that the connections among the groups grow in the last phase.

Interestingly, the results for the subnetwork induced by FAVOR and AGAINST

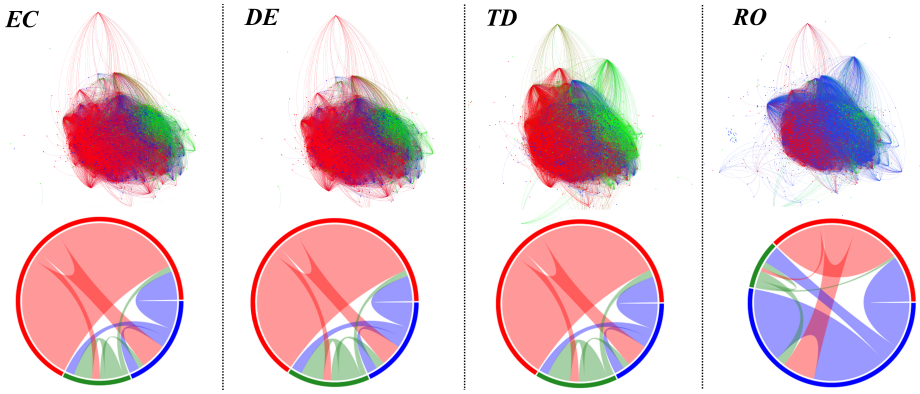

Figure 6.7: Networks based on *friend* relations for each temporal phase.

users reveal a strong homophily by stance (the rate of CE_{AF} of 0.092$\pm\sigma$0.006 is significantly lower then $2AF$ 0.324$\pm\sigma$0.013). No significant differences appear considering the four temporal phases, meaning that FAVOR and AGAINST users tend not to follow each other.

Finally, we analyzed the polarization among the three clusters computing the modularity Q_{AFN} for each temporal phase. As showed in Figure 6.5(a), the value changes during the debate starting from the minimum measured value of Q_{AFN} 0.096 after the election outcome on RO phase at a maximum of Q_{AFN} 0.164 and Q 0.160 forthcoming the election respectively on the DE and TD phases, revealing a considerable level of polarization. As observed for homophily, considering the subnetwork of AGAINST and FAVOR users, the levels of polarization are higher.

Retweets

Figure 6.8 shows the evolution of the retweets-based network along the four temporal phases. In this case, both the network structure and the users' stance may change.

As it happens considering the friend relations, the number of users labelled as NONE significantly increases in time, specially in the last phase, and again, the users affected by this phenomenon are those located in the middle of the network, connected with both the AGAINST and the FAVOR clusters as confirmed by the chord diagram.

The network exhibits a quite strong homophily considering AGAINST, FAVOR, and NONE clusters (the fraction of CE_{AFN} 0.243$\pm$0.086σ is significantly less than $2(AF + AN + NF)$ 0.344 $\pm$0.038σ). Interestingly, as in the friends-based network, an inverse trend appears in the phase RO as showed in Figure 6.4(b). Moreover, the subnetwork of AGAINST and FAVOR clusters exhibits a strong homophily by stance (the fraction of CE_{AF} 0.032$\pm$0.006σ is significantly less than $2AF$ 0.324 $\pm$0.013σ).

The retweet networks appear to be highly segregated between supporters and critics of the reform. We computed the modularities Q_{AFN} and Q_{AF} for each temporal phase as showed in Figure 6.5(b). The values change during the debate starting from the minimum measured value of Q_{AFN} 0.167 on the EC phase at a

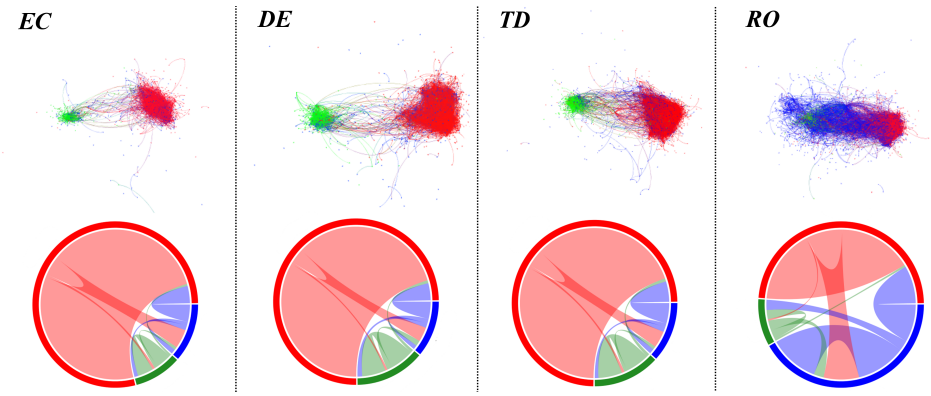

Figure 6.8: Networks based on retweet relations for each temporal phase.

maximum of Q_{AFN} 0.232 forthcoming the election on the DE phase, revealing a quite high polarization. No significant difference is observed considering the Q_{AF} values. Observe that the inverse homophily and lower values of modularity in the last phase suggest that different groups raised the number of instances of cross groups communication, but this phenomenon affects this network to a lesser extent compared to the friends-based ones. This is probably due to the fact that also communications among NONE users grow in the last phase, as it can be seen in the respective chord diagram.

Quotes

Figure 6.9 shows the evolution of the quotes-based network along the four temporal phases. Both the network structure and the users' stance are subject to change.

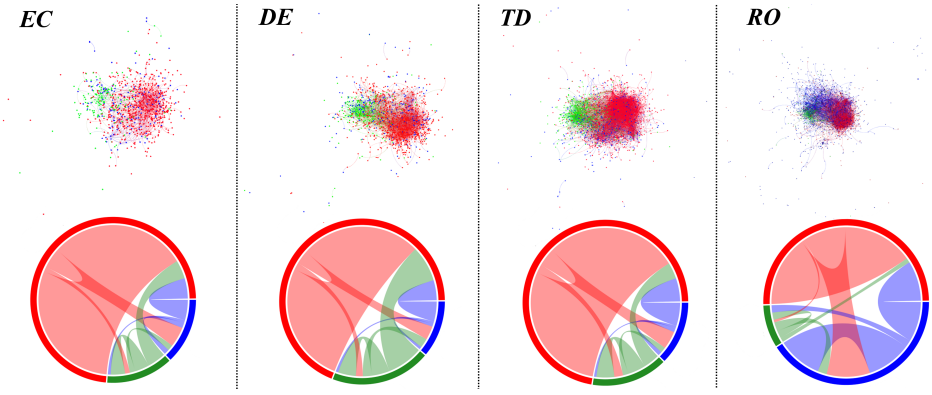

Figure 6.9: Networks based on quote relations for each temporal phase.

Once again, the number of NONE users increases over time, specially in the last phase, when a significant group of these users appears among the division between AGAINST and FAVOR clusters.

The three clusters exhibit a very light signal of homophily by stance (the fraction of CE_{AFN} $0.31 \pm 0.038\sigma$ is slightly smaller than $2(AF + AN + NF)$ $0.344 \pm 0.038\sigma$). As in the friends-based and retweets-based networks, an inverse homophily signal emerges in the phase RO. As showed in 6.4(c), a strong homophily signal is observed if we consider the subnetwork of AGAINST and FAVOR clusters (the fraction of CE_{AF} 0.106 ± 0.036 is significantly less then $2AF$ 0.324 ± 0.013).

The values of modularity change during the debate: Q_{AFN} and Q_{AF} are very similar in EC, DE, and TD phases, revealing some polarization, while they diverge in the last phase RO (see Figure 6.5(c)). In particular, there is an increasing level of polarization considering the three clusters (AGAINST, FAVOR, and NONE) and a decreasing level of polarization considering just AGAINST and FAVOR clusters in the phase RO. Observe that also in the first three phases the modularity values are smaller than the respective ones in the retweets-based networks, and this is probably due to an increment of the connections among the AGAINST and FAVOR users, as it can be seen in the chord diagrams in Figure 6.9. In particular, compared to retweets and friends based networks, there is an important high number of FAVOR users that quote AGAINST users. In the last phase this effect is mitigated in the whole network because NONE users quote are very likely to quote each other, and the three communities appear more polarized.

Replies

Figure 6.10 shows the evolution of the replies-based network along the four temporal phases. Both the network structure and the users' stance are subject to change.

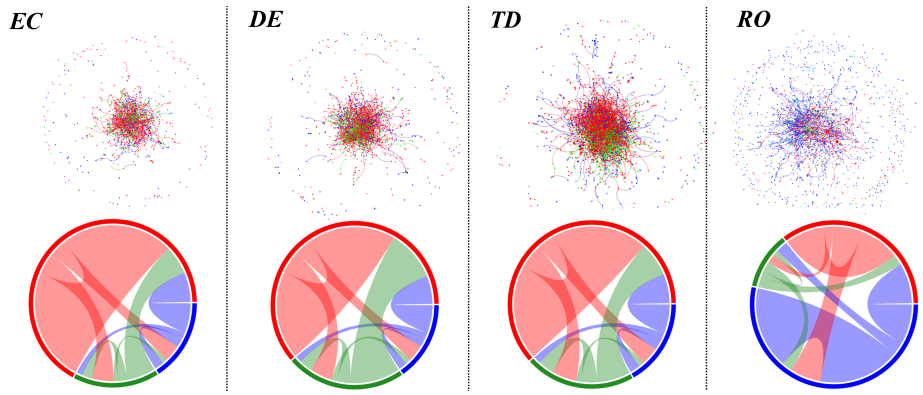

Figure 6.10: Networks based on reply relations for each temporal phase.

The replies-based networks exhibit a signal of inverse homophily by stance (the fraction of CE_{AFN} $0.443 \pm 0.053\sigma$ is significanlty higher than $2(AF + AN + NF)$ $0.344 \pm 0.038\sigma$). Moreover, the subnetworks formed by AGAINST and FAVOR clusters do not exhibit homophily by stance (the fraction of CE_{AF} $0.321 \pm 0.052\sigma$ is comparable to $2(AF)$ $0.324 \pm 0.013\sigma$). Furthermore, Figure 6.4(d) shows that the homophily values significantly change during the four temporal phases.

In this case, we do not observe striking divisions within the network as happened with the other types of relations. We computed the modularities Q_{AFN} and Q_{AF} for each temporal phase. The values change during the debate from a minimum

measured value of Q_{AFN} and Q_{AF} (respectively 0.057 and 0.024) in DE phase to a maximum of Q_{AFN} and Q_{AF} (respectively 0.166 and 0.113) in the RO phase. Figure 6.5(d) reveals a lower polarization compared to the one observed in the other types of network. This can also be seen in the chord diagrams in Figure 6.10, in which we observe a considerable number of links among different groups. However, after the second phase the polarization levels increase, meaning that cross-stance connections decrease, and this is also evident in the chord diagrams (specially in the last one).

Users' Stances Trends

We observed that users labeled with the stance NONE tend to increase in time, in particular after the referendum outcome (i.e., the RO phase). Therefore, we aim at investigating if the tendency of users to change towards the stance NONE depends on the fraction of cross-stance edges observed in the previous phase, meaning that a larger diversity of opinions can influence the propensity to change personal opinion. We computed the fraction of cross-stance edges for each user in the phases EC, DE, and TD. Then, we computed the fraction of users that change stance from the label AGAINST or FAVOR to the label NONE respectively in the following phases DE, TD, and RO. Figure 6.6 shows the relation between the fraction of cross-stance edges and the likeliness to change from AGAINST or FAVOR to NONE for each network type (friends, retweets, quotes, and replies networks). The dashed lines are linear polynomials that interpolate the discrete set of known data points. The percentage of users that changes from AGAINST or FAVOR to NONE is not negligible (about 16%).

The results confirm that users with more heterogeneous connections are also more likely to change their annotated stance to NONE. In particular, users who exhibit a high rate of cross-stance friend relationships, tend to conceal their opinion more easily in the next phase. This happens in each type of relations on which the network is based. In the replies-based network, we already observed that the users tend to create a higher number of heterogeneous connections compared to the other networks. Nevertheless, Figure 6.6(d) shows smaller probabilities to change opinion and a smaller dependence on the number of cross-stance connections, compared to other relations-based graphs. Apparently, fighting for defending opinions, reinforce the beliefs themselves, leading to a reduced probability of changing stance.

6.5.4 Discussion

In this section we analyzed a manually annotated Italian corpus (i.e., *ConRef-STANCE-ita*) for addressing stance detection from a diachronic perspective, contributing to understand social networks and opinions dynamics. We observed that, in this particular framework (the Twitter debate about the 2016 Italian Constitutional Referendum), an increasing fraction of users tend to conceal their opinion, especially after the outcome. Indeed, a significant number of users previously labeled with a clear stance (FAVOR or AGAINST), have been labeled with the stance NONE in a following phase of the debate, suggesting that users' stances are less explicit, therefore the annotators were no longer able to infer their opinion.

The investigation of network structures led to the observation that users are generally aggregated in homogeneous communities, except for the replies-based network. This is reasonable since, users having different opinions often tend to discuss

using replies [25]. We showed that the network structures based on friends, retweets, and quotes exhibit a clear homophily by stance among supporters and critics of the reform, suggesting that users tend to connect more likely to others with the same opinion. However, an inverse homophily by stance emerges in the last phase of the debate for all these types of networks; in other words, in the last phase there are more connections among users labeled with different stances. Nevertheless, the replies-based network shows an inverse homophily by stance during the whole debate, suggesting that "reply to" communication instances are preferred when a discussion on different positions in the debate arises between two users. Also the modularity values reveal quite high levels of polarization in friends, retweets, and quotes based networks and an increasing polarization appears in the replies-based network forthcoming the elections and after the outcome: apparently, maintained discussions between users with different opinions just augmented distances instead of reducing them.

Finally, since the number of NONE labeled users increases during the debate, we explored the relation between the level of diversity in the neighborhood of FAVOR and AGAINST users (number of cross-stance edges) and the likelihood to be labeled as NONE in the next phase. Results suggest that users who exhibit a higher fraction of cross-stance connections tend to conceal their stance more frequently in the following phase of the debate.

6.6 Talent Identification as a Binary Classification Task

This thesis mainly focuses on stance detection and on the exploration of features extracted from the network structure of the users' participating in polarized debates. In this section, on the basis of some of the results we published in Lai et al. [44], we aim at potentially extending our contribution to other issues that could be addressed as binary classification tasks taking advantage of features extracted from a graph representation of the problem. Indeed, a lot of problems could be represented as a graph where entities are nodes and relations among entities are edges. Our hypothesis is that network based features can be successfully used in other tasks different from stance detection. To corroborate our hypothesis, we tried to address the task of talent identification taking advantage of network-based features. The description of the problem and the conducted experiments are described in this section.

6.6.1 Case Study

Here, we address the problem of talent identification in sport particularly focusing on the case study of table tennis. In this task we focused on the identification of talents, where a talent is defined as a young athlete with age between 8 and 14 that reaches the top 10% of the national rank at the end of the season. We addressed the problem as a binary classification task where a label “talent” or “not talent” is predicted for each athlete. We trained the model with the information of the athletes on the season S and we predict the talented athletes at the end of the season $S + 1$.

A large dataset containing all table tennis matches played in Italy between the five seasons from 2011/12 to 2015/16 has been released by the Italian Table

Tennis Federation (FITET[17]) and has been used for performing the experiments. In particular, the dataset consists in 723,057 table tennis matches played by 21,458 players. We create a graph for each season where nodes are table tennis players and the matches played between two players are the edges. The athletes with age between 8 and 14 were 2916, and 230 were considered talented. The dataset is strongly biased toward the label "not talent", so we focus on maximizing the precision and the recall of the class "talent".

6.6.2 Methodology

We exploited two groups of features that involve *demographics*, *performance* and *topological network based* attributes. Following, we describe these features.

Demographics

Attempting to take advantage of players' population statistics information, two features were included in this group:

- Age: The age of the athlete at the season S. Figure 6.11 shows the label distribution over the different ages.

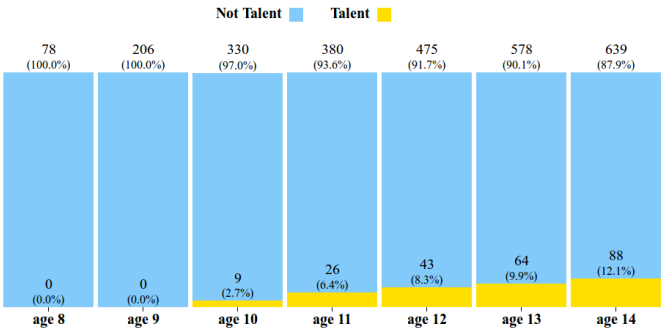

Figure 6.11: Label distribution of labels "not talent" and "talent" over the different ages.

- Gender: The gender of the athlete. 2071 athletes are male and 845 female. The talents are 194 male and 36 female.

Performance

This group of features explores the sport performance of the athlete in the season S.

- Matches: Number of matches played
- Won matches: Number of won matches

[17] www.fitet.org

- Lost matches: Number of lost matches
- Won Sets: Number of won sets
- Lost Sets: Number of lost sets
- Sets difference: Difference between the number of won and lost sets

---

Topological Network Based Features

Sport performance studies usually focus on training activities, nutrition, behavior, and cognitive strategies. But several experts agree that self-improvement directly depends on the competitions and the opponents a player is exposed to. In other words, to play with stronger opponents that use diverse styles of play improves athlete's experience and competitiveness. Starting from these intuitions, we try to analyze how the topological properties of the graph of matches could help for identifying talent. We create an directed graph where athletes are nodes and and edges between to athletes exists if they played against each other (a graph for each season between 2011/12 to 2015/16 was created). In this way we discard any information on the performance of a player (e.g. the number of won matches) exclusively maintaining the connectivity patterns. The following features are extracted calculating the network metrics from the network of the seasons S:

- **Centrality**
  - Eigenvector Centrality: It measures the influence of a node in a network computing the centrality of a node by the centrality of its neighbors.
  - Load Centrality: It measures the fraction of all shortest paths that pass through that node.
  - Betweenness Centrality: It measures the number of shortest paths that pass through a node; it is similar to load centrality.
  - Degree Centrality: It is a centrality indicator that identifies the most important nodes on the basis of the node degree.
  - Closeness Centrality: It measures the inverse of the sum of the shortest distances between each node and every other node in the network.
- **Page Rank** The measure computes a ranking of the nodes in the graph G based on the structure of the in-edges. The network is undirected, thus each edge is converted in the two directed edges.
- **Average Neighbor Degree** It computes the average degree of the neighborhood of each node.

6.6.3 Experiments

We addressed the talent identification task as a binary classification problem exploiting linear support vector machine (SVM), random forest (RF), and logistic regression (LR) as machine learning algorithms. We carried out several experiments by combining the proposed groups of features performing a 5-cross validation.

In the first experiment, we create a directed graph where athletes are nodes and a directed edges between A and B exists when B defeated A (a graph for each season between 2011/12 to 2015/16 was created). The results are showed in Table 6.22.

Table 6.22: Precision and recall using SVM trained with the different groups of features

	LR		**RF**		**SVM**	
Group of Features	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
Demographics	0	0	0	0	0	0
Performance	0.769	0.609	0.648	0.639	0.793	0.583
Topological	0.625	0.326	0.803	0.796	**0.818**	0.156
Demographics+ Performance	0.781	0.622	0.668	0.604	0.794	0.604
Demographics+ Topological	0.694	0.404	0.794	0.756	0.758	0.409
Performance+ Topological	0.776	0.648	0.836	0.778	0.799	0.604
Demographics+ Performance+ Topological	**0.781**	**0.665**	**0.839**	**0.791**	0.816	**0.674**

The *demographics* features are not able to predict talent and the feature vector consisting in the two elements Age and Gender is useless, unless used together with the other group of features. Instead, the *performance* features obtain good results and reach the highest value of precision using SVM and the highest value of recall using RF. It is reasonable to assume that an athlete that shows high performance in season S could begin a talent candidate for the future season $S+1$. For this reason, we consider the group of *performance* features as a sort of baseline. The group of *topological network based* features obtains the highest precision using SVM (0.818), but a very low recall (0.156). The best result is obtained using RF (f-measure of 0.799 is achieved with a precision of 0.803 and a recall of 0.796). Notably, the result obtained from the group of *topological network based* features using RF improves the one achieved with the *performance* features. Furthermore, *demographics*, *performance*, and *topological network based* features used together overcome the results of the other combination of features. In particular, the best precision and recall are obtained using RF (respectively 0.839 and 0.791).

In the second experiment, we create an undirected graph where athletes are nodes and an edge between to athletes exists if they played against each other. In this way we discard any information on the performance of a player (e.g. the number of won matches) from the graph and we exclusively maintain the connectivity patterns from the *topological network based* features.

The results are showed in Table 6.23 Only the *topological network based* features have been modified, thus only the combination of the group of features containing them are showed.

The results achieved by the *topological network based* features considerably decrease, unless used together with the other group of features. In particular, are useless for training SVM. The result obtained by the combination of *performance* and *topological network based* features reaches the highest value of f-measure with

Table 6.23: Precision and recall using SVM trained with the different groups of features using an unweighted network

	LR		**RF**		**SVM**	
Group of Features	*precision*	*recall*	*precision*	*recall*	*precision*	*recall*
Topological	0.3125	0.043	0.558	0.504	0	0
Demographics+ Topological	0.429	0.0529	0.719	0.635	0	0
Performance+ Topological	**0.793**	**0.617**	0.618	0.535	**0.815**	0.613
Demographics+ Performance+ Topological	0.786	0.609	**0.745**	**0.648**	0.796	**0.661**

LR (0.694 with 0.793 of precision and 0.617 recall) despite the discarding of any information on the performance of a player from the network structure. Summarazing, this means that the feature based on the network structure captures relevant information for discriminating talents without any information of the performance of the athlete.

6.6.4 Discussion and Conclusion

In this section we showed that modeling a problem as a graph helps in addressing a different binary classification task. Here, we captured relevant information from the network of tennis tables athletes for predicting talents. We started from the intuition that the opponents contribute to the development of a player's career and we used matches as edges for connecting two athletes for identifying the most important nodes (talents) within a graph of athletes. Results showed that combining *demographics*, *performance* and *topological network based* features gives consistently the best performance in this task. These observations have potentially a strong impact on defining innovative methodologies to model other classification tasks focusing on network analysis. With this case study, we showed that extracting features from a network structure is not a procedure exclusively useful for stance detection, but, on the contrary, that this technique could applicable in other classification tasks.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This thesis describes our approach to the problem of stance detection in Twitter. Focusing on political polarized debates, we explored contextual features for proposing new models able to improve stance prediction. Our research also allows us for inspecting the dynamics of communication among people having concordant or contrasting opinions, with a particular focus on observing opinion shifting.

We addressed stance detection as a binary classification task taking advantage of support vector machine, logistic regression, and naive Bayes machine learning algorithms. Part of our contributions relies on proposing an extensive set of textual and contextual features. The first contextual feature we proposed is based on relations among the target of interest and its friends or enemies, i.e. DOMAIN KNOWLEDGE (Chapter 2). We tested the feature in a benchmark dataset for English [60] showing that it is very useful when the target is not mentioned in the text. Indeed, the feature is able to predict the stance towards the target capturing the opinion expressed toward a related entity. We proposed and tested two new contextual features, i.e LANGUAGE and URL, on a second benchmark dataset [91] proposing the model iTACOS in a shared task on stance and gender detection in Catalan and Spanish (Chapter 3). An ablation and an additive test on iTACOS allowed us to really appreciate as contextual features played a part in obtaining the higher results for stance detection; the model ranks first in both Catalan and Spanish sub-tasks (Section 6.2). Furthermore, we showed that, performing a cross-validation over the training set, the model achieves highest results compared to those obtained on the test set. As a general observation statistical approaches could create a built-in model based on the specific dataset used for training it. In this case, probably an over-fitting occurred. For this reason, we should seek to only propose features that realistically are consistent with the problem we address, although some features seem to obtain high f-measures performing experiments on the training set.

Then, in Section 6.3, we addressed stance detection in a multilingual perspective for investigating the portability of iTACOS, here called MULTITACOS, across different languages. We explored two benchmark datasets and we created two novel datasets to further explore stance detection in a multilingual perspective. The English benchmark dataset [60] E-USA focuses on two targets related to political elections in the USA ("Hillary Clinton" and "Donald Trump"), thus we collected

tweets about the two candidates of the run-off of the French presidential election held in 2017 i.e., "Emmanuel Macron" and "Marine Le Pen" for the French dataset E-FRA. For what concerns instead the Italian dataset R-ITA, we collected tweets for Italian about the "Constitutional Referendum" held in 2016 in Italy, which mirrors with the target of the benchmark Spanish-Catalan corpus "Independence of Catalonia" [91] that we referred to R*-CAT. We proposed four groups of features for addressing stance detection: *Stylistic*, *Structural*, *Affective*, and *Contextual*. *Stylistic*. *Structural* features obtained good results in all supervised contexts and lower results in semi-supervised ones. Indeed, similarly to what we observed testing iTACOS, bag of tokens such as Bag of Char-grams and Bag of Twitter Marks (hashtags and mentions) significantly contribute to obtain high results in supervised contexts. On the other hand, in semi-supervised contexts, the best results are obtained using models which exploit *Affective* and *Contextual* features. With respect to *contextual* features, we showed that the feature domain knowledge is more relevant in supervised contexts when the target is a person. Then, the feature url takes advantage from the fact that users normally tend to share links to web pages which support their position (e.g. towards the "Italian Constitutional Referendum"). Finally, the language feature is particularly discriminating with the target "Independence of Catalonia" where nationalist feelings play a big role and language itself is exploited to convey Catalan independentist attitude.

Subsequently to, our research shows that knowing the social network community an user belongs, helps in detecting her stance towards a particular target of interest. We also inspected the dynamics of communication among people having concordant or contrasting opinions particularly focusing on opinion shifting. Benchmark datasets do not allow to explore users' context due to they usually only consists in a set of annotated texts without considering any information about the user. For this reason, we created TW-BREXIT, a corpus of English annotated tweets for stance towards the European Union membership referendum (Chapter 4). From a user perspective, we splitted the debate in three 24-hours temporal phases and we selected a random sample of users that wrote at least 3 tweets in each time interval. This allows us to know the stance towards BREXIT of all users and to observe opinion shifting. In particular, the evidence suggests that users tend to conceal their stance approaching to the referendum outcome and that they tend to create within-stance relations building segregated communities. Furthermore, using the communities extracted from the network of friendship relations among users, helps for detecting users' stances. In Section 6.4, we employed features such as Bag of Word, structural, sentiment, and contextual based features for predicting stance in TW-BREXIT. Support vector machine obtains the best result taking advantage of Bag of Word, sentiment-based features, diachronic-evolution-context-based features, and community-context-based features. The community-context-based feature appears to be of particular importance for stance detection in ablation tests. Furthermore, this feature achieves very relevant results training tree models such as Decision Trees and Random Forests.

We also reported a similar case of study analyzing the debate about another referendum i.e., the Italian Constitutional Referendum. We created *ConRef-STANCE-it*, a corpus of annotated Italian tweets for stance detection towards the target "Italian Constitutional Referendum" [45] (Chapter 5). We splitted the debate in four 72-hours temporal phases and we selected a random sample of users that wrote at least one tweet, one retweet, and one reply in each time interval. It allows us for looking more closely at opinion shifting taking advantage of a wider time windows.

Furthermore, requiring a retweet and a reply for each user in each time window, allows us for extending our analysis on networks based on retweets and replies. Due to a significant number of quote relations among users involved in the debate, we analyzed also the network based on quotes.

In Chapter 5, we performed classification experiments using features based on the textual content of the tweet and on the community the author belongs to extracting information from the networks based on retweets, replies, and quotes. The features extracted from retweets and quotes based networks considerably improve the accuracy of stance detection than when using only the the features based on the textual content. This does not happen using the feature extracted from the replies-based networks. Indeed, although the users mainly reply to other users with a similar opinion, we observe about 20% of cross-stance edges. This is a particularly interesting case where inverse homophily (or also heterophily) could be observed.

In Section 6.5, we further carried out an analysis on this case of study focusing on the dynamics of communication among people having concordant or contrasting opinions, particularly focusing on observing opinion shifting. First, we observed an evidence also reported in TW-BREXIT: a significant number of users previously labeled with a clear stance (FAVOR or AGAINST) have been labeled with the stance NONE in a following phase of the debate. This suggests that users tend to conceal their stance; therefore the annotators were no longer able to infer their opinion. Similarly, the users are generally aggregated in homogeneous communities, except for the replies-based network. This evidence suggests that users tend to connect more likely to others with the same opinion. The network structures based on friends, retweets, and quotes exhibit a clear homophily by stance among supporters and critics of the reform. However, there are more connections among users labeled with different stances in the last phase. Indeed, an inverse homophily by stance emerges in the last phase of the debate for all these types of networks. Nevertheless, an inverse homophily by stance has been observed in the replies-based network during the whole debate. It suggests that users prefer to use replies instead of other communication types when they have a confrontation with different point of view. Moreover, the levels of modularity reveal a quite high polarization in friends, retweets, and quotes based networks. The polarization appears to increase in the replies-based network forthcoming the elections and after the vote outcome. Furthermore, we observed that the likelihood to be labeled as NONE in the next phase of the debate depends on the fraction of cross-stance connections. Indeed, more cross-stance relations exist, more likely the user tends to be less explicit in the following phase of the debate.

Finally, we explored a different case of study with the aim of verifying that our method, in particular the extraction of the features from a network representation of the problem, is not limited to stance detection, but can be applied to different scenarios. To demonstrate that, we faced the problem of talent identification in sport particularly focusing on the case of study of table tennis. We showed that network metrics based on centrality capture relevant information from the network of tennis tables athletes that cloud be used by a machine learning algorithm for predicting talent.

The results described in this thesis have attempted to answer the research questions we introduced in the Chapter 1:

i) Is contextual information useful for SD in social media?

Yes. Large part of this thesis has been focused on describing the contribute of contextual features for detection stance. We explored the features based on *domain knowledge*, *language* and *url* and we showed the contributions they can bring to stance detection in different case of studies. Particularly, we highlighted the high contribution that the feature based on the *community* the author of the tweet belongs to may have.

ii) *Among the contextual features, what is the impact of the network structure?*

We showed that the network structure captures relevant information for stance detection. Indeed, probably due to the fact that users tend to bond with like-minded, the knowing of the community an user belongs to is a strong signal for stance detection. We explored two case of studies in Chapters 4 and 5 following up on the problem in Sections 6.4 and 6.5.

iii) *How to address SD from a multilingual perspective?*

Analyzing four different political debates in five different languages such as English, Spanish, Catalan, French, and Italian (Section 6.3), we showed that the proposed method achieves good results in all languages. On the other hand, we reported that the features we have to use for training the model depend on the typology of the target (person or election/referendum).

iv) *Could the feature based on social network structure be used in other classification tasks?*

We proposed another classification task that performs talent identification in sport, particularly focusing on the case of study of table tennis. We showed that networks metrics based on centrality are strong signal for talent and can be used for training a machine learning algorithm model for this task (Section 6.6).

Analyzing the debate focusing on both users' stance and relations in a diachronic perspective, we finally answered the following questions (Chapters 4, 5 and Sections 6.4 and 6.5):

v) *Are there benefits for addressing SD from a diachronic perspective?*

To a certain extent, yes. Indeed, we showed that the label distribution changes over the time even considering the same sample of users. Therefore, the annotated stance of an user is valid for a short interval of time and we need to regularly "update" it. Furthermore, we know that the label distribution of a corpus influences the training of a model, particularly in machine learning algorithms based on probabilistic cues. For these reasons, we have to take into account the temporal context of the political debate.

vi) *Could stance label distribution change during time?*

Yes. We showed that label distribution changes over the time in two case of studies (BREXIT and Italian Constitutional Referendum political debates, Chapters 4 and 5). In particular, we showed that the distribution of the

stance NONE increases approaching to the referendum outcome. Probably, the annotator is more difficult to infer users' stance due to the fact that they are less explicit expressing their stance when the debate approaches to a conclusion. Further studies on the temporal evolution of communication styles in political debates are needed for shedding more light on these hypothesis.

vii) *Could the stance of the tweeter influence the type of relation s/he creates with other users?*

Yes. Although, replies are also extensively used for establishing within-stance relations. In Section 6.5, we showed that users prefer to use replies for communicating with others having different point of views. Quotes and retweets, on the contrary, are almost exclusively used for creating within-stance relations.

viii) *Could the neighbours influence the probability of observing an opinion shifting?*

Yes. Indeed, as we showed, in Sections 6.4 and 6.5, it is more likely that a user changes her stance to NONE, in the case of having a high number of cross-stance relations in the previous phase of the debate.

7.2 Research Contributions

We outline our publications during the Ph.D. by grouping them into four main groups.

1) Stance detection using contextual features

We described our approach for predicting stance taking advantage of contextual features in two conference papers:

- **Lai M.**, Hernández Farías D.I., Patti V., Rosso P. (2017) Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweet. In: *Proceedings of the 15th Mexican International Conference on Artificial Intelligence (MICAI 2016)*. Cancún, Mexico, October, 23–28, 2016. Lecture Notes in Computer Science, vol 10061, pages 155-168. Springer.
  ISBN: 978-3-319-62433-4
  DOI: https://doi.org/10.1007/978-3-319-62434-1_13
- **Lai, M.**, Cignarella A.T., Hernández Farías D.I. (2017) ITACOS at ibereval2017: Detecting stance in Catalan and Spanish tweets. In: *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*. Murcia, Spain, September 19, 2017. CEUR Workshohp Proceedings, vol. 1881, pages 185–192. CEUR-WS.
  ISSN:1613-0073

2) Exploring political polarized communities

We inspected users' online social network for proposing contextual features based on the network structure. We presented our results to the research community in two conference papers.

– **Lai M.**, Tambuscio M., Patti V., Ruffo G., Rosso P. (2017) Extracting Graph Topological Information and Users' Opinion. In: *Proceedings of the Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017)*. Dublin, Ireland, September 11-14 2017. Lecture Notes in Computer Science, vol. 10456, pages 112-118. Springer.
ISBN: 978-3-319-65812-4
DOI: https://doi.org/10.1007/978-3-319-65813-1_10

– **Lai M.**, Patti V., Ruffo G., Rosso P. (2018) Stance Evolution and Twitter Interactions in an Italian Political Debate. In: *Proceedings of the 23rd International Conference on Natural Language & Information Systems (NLDB 2018)*. Paris, France, June 13-15 2018. Lecture Notes in Computer Science, vol. 10859, pages 15-27. Springer.
ISBN: 978-3-319-91946-1
DOI: https://doi.org/10.1007/978-3-319-91947-8_2

3) Applying network metrics for binary classification in other machine learning tasks

We explore social community for proposing features for other machine learning tasks in a journal paper showing that our approach may be useful in other domains.

– **Lai M.**, Meo R., Schifanella R., Sulis E. (2018) The role of the network of matches on predicting success in table tennis. In: *Journal of Sports Sciences*.
ISSN: 0264-0414
DOI: https://doi.org/10.1080/02640414.2018.1482813

4) Other contributions

Below, a list of additional research works we presented to the scientific community during this PhD..

First, we list the papers that address political debates on Twitter. They are partially related to the linguistic analysis of stance and irony detection.

– **Lai M.**, Bosco C.; Patti V., Virone D. (2015) Debate on Political Reforms in Twitter: A Hashtag-driven Analysis of Political Polarization. In: *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA 2015)*. Paris, France, October 19-21, 2015. 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pages 1-9. IEEE.
ISBN: 978-1-4673-8272-4
DOI: https://doi.org/10.1109/DSAA.2015.7344884

– Virone D., **Lai M.** (2015) Dans un corpus hybride : les messages twittés, l'intertextualité et la formule. In: SHS Web of Conferences, 20, 01021
DOI: https://doi.org/10.1051/shsconf/20152001021

– **Lai M.**, Virone D.; Bosco C., Patti V. (2015) Building a Corpus on a Debate on Political Reform in Twitter. In: *Proceedings of the Second Italian Conference on Computational Linguistics (CLiC-it 2015)*. Trento, Italy, December 3-4, 2015. Collana dell'Associazione Italiana

di Linguistica Computazionale, pages 171,176. aAcademia University Press.
ISBN: 978-88-99200-62-6

- Bosco C., **Lai M.**, Patti V. and Virone D. (2016) Tweeting and Being Ironic in the Debate about a Political Reform: the French Annotated Corpus Twitter-MariagePourTous. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia, May 23-28, 2016. Pages 1619-1626. European Language Resources Association (ELRA).
ISBN: 978-2-9517408-9-1
- Bosco C., **Lai M.**, Patti V., Rangel Pardo F. M., Rosso P. (2016) Tweeting in the Debate about Catalan Elections. In: *Proceedings of the LREC 2016 Workshop "Emotion and Sentiment Analysis" (LREC 2016)*. Portorož, Slovenia, May 23, 2016. Pages 67-70. European Language Resources Association (ELRA).

The conference paper below describes the creation of an Italian corpus for irony detection in specific polarized debate in different languages (French, Italian, Spanish and Catalan).

- Cignarella A. T., Bosco C., Patti V., **Lai M.** (2018) Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan, May 7-12, 2018. Pages 4204-4211. European Language Resources Association (ELRA).
ISIN: 979-10-95546-00-9

Following we list two conference papers that integrate official statistics and social media data taking advance of sentiment analysis techniques.

- Sulis M., Bosco C., Patti V., **Lai M.**, Farías Hernández D. I., Mencarini L., Mozzachiodi M, Vignoli D. (2016) Subjective Well-Being and Social Media: A Semantically Annotated Twitter Corpus on Fertility and Parenthood In: *Proceedings of the 3rd Italian Conference on Computational Linguistics (CLiC-it 2016) and the 5th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2016)*. Naples, Italy, December 5-7, 2016. CEUR Workshohp Proceedings, vol. 1749. CEUR-WS.
ISSN: 1613-0073
- Sulis M., **Lai M.**, Vinai M.,Sangunetti M. (2015) Exploring sentiment in social media and official statistics: A general framework In: *Proceedings of the 2nd International Workshop on Emotion and Sentiment in Social and Expressive Media: Opportunities and Challenges for Emotion-aware Multiagent Systems (ESSEM 2015) co-located with 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*. Istanbul, Turkey, May 5, 2015. CEUR Workshohp Proceedings, vol. 1351, pages 96-105. CEUR-WS.
ISSN: 1613-0073

Finally, a journal paper that evaluates human-computer interaction in a in self-tracking case of study.

- Rapp A., Marcengo A., Buriano L., Ruffo G., **Lai M.**, Cena F. (2018) Designing a personal informatics system for users without experience in self-tracking: a case study In: *Behaviour & Information Technology*, 37:4, pages 335-366
  DOI: https://doi.org/10.1080/0144929X.2018.1436592

7.3 Future Work

Stance detection is a relatively new computational linguistic task that is rapidly gaining exposure in the research community. State-of-the-art approaches achieve results that are not far from those obtained by the baselines proposed in the two shared tasks [60, 91]. That means that there is still plenty to do for improving stance detection performances. Following, we mention some areas that could be investigated for addressing this issue in further studies:

- Domain knowledge is also needed for human annotators for inferring users' stance. We showed that to know the relations among the target of interest and its friends or enemies helps in approaching the task (Chapter 2). We really should spend some time increasing the number of relations between the target of interest and other entities taking advantage of domain experts. We could also try to automatically gather the relations from online ontologies such as Dbpedia. For politicians the properties `dbp:children` or `dbo:spouse` could by useful for retrieving personal information and the property `dbo:party` (`is dbo:party of`) could help in retrieving the party affiliation and the list of party colleagues.
- Users often use irony in order to express their stance towards the target of interest. Although there are many works focusing on the fact that irony is often used for inverting the polarity, which is a problem in sentiment analysis (polarity reversal) [88], no work specifically focuses on the role of irony in stance detection. We believe that to combine stance and irony detection could be an interesting future research line to improve stance detection.
- Conversational thread are analyzed for detecting the type of interaction between a given tweet and its reply (SemEval-2017 Task 8 [18]). Stance detection could be also useful in determining rumour veracity and support for rumours detecting the user's stance towards the rumor. We can also explore the type of interactions within and across communities.

We also shed some light in users' behavior in social media investigating online relations among agreeing and conflicting point of views. In several research areas it could be interesting to investigate group formation and segregation in social media taking advantage of machine learning approaches:

- We reported that the distribution of the label NONE increases as the date of the referendum approaches. Something changes in the way the users express their stance to the point that the annotators are no longer able to infer their opinion. A linguistic analysis for inspecting what changes in communication strategies could be useful for shedding more light on opining shifting.

- Social media allow for inspecting group formation and group polarization. An deeper analysis of the different communication strategies used within and across communities could help in observing and in preventing the formation of extremist view points.
- We also propose to detect ironical intents and to verify if different types of irony are exploited. In particular, we are interested in verifying if irony is used for joking adversary within community and for offending opponents in communication across communities (sarcasm).

We finally showed as network structure helps to improve stance detection. We also show as this approach can be used in very different tasks such as talent and performance prediction. Therefore, we believe that features based on the network structure should be exploited also in other tasks. Following, we provide some examples of potential applications:

- We hypothesize that irony and sarcasm could respectively prevail in communications within and across communities, respectively. For this reason, we believe that community detection could provide a strong cue for detecting these types of figurative messages.
- Community detection could also help in detecting fake news. Indeed, sources of fake news strengthen their credibility mentioning each other. For this reason, a news could provide a strong signal of fake news if the author belongs to a community of unverified sources.
- In general network analysis could help also for author profiling (e.g. age and gender), sexual predators detection, hate speech detection, and other tasks providing a community context of the author in support of the textual cues extracted from the analyzed post.

Bibliography

[1] Lada A. Adamic and Natalie Glance (2005), The political blogosphere and the 2004 u.s. election: Divided they blog. In: *Proceedings of the 3rd International Workshop on Link Discovery*, 21-24 August 2005, Chicago, Illinois (LinkKDD 2005), pp. 36–43. Association for Computing Machinery, New York, NY, USA.

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau (2011), Sentiment analysis of Twitter data. In: *Proceedings of the Workshop on Languages in Social Media*, 23 - 23 June 2011, Portland, Oregon (LSM 2011), pp. 30–38. Association for Computational Linguistics, Stroudsburg, PA, USA.

[3] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor (2011), Cats rule and dogs drool!: Classifying stance in online debate. In: *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 24 June 2011, Portland, Oregon (WASSA 2011), pp. 1–9. Association for Computational Linguistics, Stroudsburg, PA, USA.

[4] Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu (2008), Distilling opinion in discourse: A preliminary study. In: *22nd International Conference on Computational Linguistics: Companion volume: Posters*, 18-22 August 2008, Manchester, UK (Coling 2008), pp. 7–10. Association for Computational Linguistics, Stroudsburg, PA, USA.

[5] Isabelle Augenstein, Andreas Vlachos, and Kalina Bontcheva (2016), USFD at SemEval-2016 task 6: Any-target stance detection on Twitter with autoencoders. In: Bethard et al. [8], pp. 389–393.

[6] Albert-László Barabási and Zoltán N. Oltvai (2004), Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113.

[7] Albert-László Barabási (2016), *Network science*. Cambridge University Press, Cambridge, UK.

[8] Steven Bethard, Daniel M. Cer, Marine Carpuat, David Jurgens, Preslav Nakov, and Torsten Zesch (2016). *Proceedings of the 10th International Workshop on Semantic Evaluation*, 16-17 June 2016, San Diego, CA, USA (SemEval 2016), Stroudsburg, PA, USA. Association for Computer Linguistics.

[9] Vincent D. Blondel, Jean-Loup Guillaume, Jean-Loup Lambiotte, and Etienne Lefebvre (2008), Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 10:10008–10020.

[10] Henrik Bøhler, Petter Asla, Erwin Marsi, and Rune Sætre (2016), IDI@NTNU at SemEval-2016 task 6: Detecting stance in tweets using shallow features and GloVe vectors for word representation. In: Bethard et al. [8], pp. 445–450.

[11] Cristina Bosco, Mirko Lai, Viviana Patti, Francisco Manuel Rangel Pardo, and Paolo Rosso (2016), Tweeting in the debate about Catalan elections. In: *LREC workshop on Emotion and Sentiment Analysis Workshop*, 23-28 May 2016, Portorož, Slovenia (LREC 2016), pp. 67–70. European Language Resources Association, Paris, France.

[12] Cristina Bosco, Mirko Lai, Viviana Patti, and Daniela Virone (2016), Tweeting and being ironic in the debate about a political reform: the French annotated corpus Twitter-mariagepourtous. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 23-28 May 2016, Portorož, Slovenia (LREC 2016), pp. 1619–1626. European Language Resources Association, Paris, France.

[13] Cristina Bosco and Viviana Patti. Social media analysis for monitoring political sentiment. In: *Encyclopedia of Social Network Analysis and Mining*, pp. 1–13. Springer, New York, NY, USA, 2017.

[14] CENSIS (2016), *50º rapporto sulla situazione sociale del paese 2016*. Franco Angeli.

[15] Tony F. Chan, Gene H. Golub, and Randall J. LeVeque. Updating formulae and a pairwise algorithm for computing sample variances. Technical Report STAN-CS-79-773, Stanford University, Stanford, CA, USA, 1979.

[16] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Goncalves, Filippo Menczer, and Alessandro Flammini (2011), Political polarization on Twitter. In: *International AAAI Conference on Web and Social Media*, 17-21 July 2011, Barcelona, Spain (ICWSM 2011), pp. 89–96. Association for the Advancement of Artificial Intelligence, Palo Alto, CA, USA.

[17] William Deitrick and Wei Hu (2013), Mutually enhancing community detection and sentiment analysis on Twitter networks. *Journal of Data Analysis and Information Processing*, 1:19–29.

[18] Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga (2017), Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 3-4 August 2017, Vancouver, Canada. pp. 69–76. Association for Computational Linguistics, Stroudsburg, PA, USA.

[19] Marcelo Dias and Karin Becker (2016), INF-UFRGS-OPINION-MINING at SemEval-2016 task 6: Automatic generation of a training corpus for unsupervised identification of stance in Tweets. In: Bethard et al. [8], pp. 378–383.

[20] David Easley and Jon Kleinberg (2010), *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, Cambridge, UK.

[21] Erick Elejalde, Leo Ferres, and Eelco Herder (2017), The nature of real and perceived bias in chilean media. In: *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 04 - 07 July 2017, Prague, Czech Republic (HT 2017), pp. 95–104. Association for Computing Machinery, New York, NY, USA.

[22] Heba Elfardy and Mona Diab (2016), CU-GWU perspective at SemEval-2016 task 6: Ideological stance detection in informal text. In: Bethard et al. [8], pp. 434–439.

[23] Ash Evans (2016), Stance and identity in Twitter hashtags. *Language@Internet*, 13(1).

[24] David Garcia, Adiya Abisheva, Simon Schweighofer, Uwe Serdült, and Frank Schweitzer (2015), Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet*, 7(1):46–79.

[25] Kiran Garimella, Ingmar Weber, and Munmun De Choudhury (2016), Quote rts on Twitter: Usage of the new feature for political discourse. In: *Proceedings of the 8th ACM Conference on Web Science*, 22-25 May 2016, Hannover, Germany (WebSci 2016), pp. 200–204. ACM, New York, NY, USA.

[26] Andrew Gelman and Gary King (1993), Why are american presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science*, 23(04):409–451.

[27] Ungeheuer Gerold and Wiegand Herbert Ernst (2008), *Handbooks of Linguistics and Communication Science*. Walter de Gruyter GmbH & Co. KG, Berlin, Germany.

[28] Marta C. González, Cesar A. Hidalgo, and Albert-Laszlo Barabasi (2008), Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

[29] Bruno Gonçalves, Nicola Perra, and Alessandro Vespignani (2011), Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PloS one*, 6(8):e22656.

[30] Palash Goyal and Emilio Ferrara (2018), Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78 – 94.

[31] Mark S. Granovetter (1973), The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380.

[32] Tero Harju (2011), *Lecture Notes on Graph Theory*. Department of Mathematics University of Turku, Turku, Finland.

[33] Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso (2016), Irony detection in Twitter: The role of affective content. *ACM Transactions on Internet Technology*, 16(3):19:1–19:24.

[34] Bernie Hogan (2010), The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of Science, Technology & Society*, 30(6):377–386.

[35] Philip N. Howard and Bence Kollanyi (2016), Bots, #strongerin, and #brexit: Computational propaganda during the uk-eu referendum. *ArXiv e-prints*.

[36] Minqing Hu and Bing Liu (2004), Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 22-25 August 2004, Seattle, WA, USA (KDD 2004), pp. 168–177. Association for Computing Machinery, New York, NY, USA.

[37] W. John Hutchins, Leon Dostert, and Paul Garvin (1955), The georgetown-ibm experiment. *Machine translation of languages*, pp. 124–135.

[38] Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui (2016), Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In: Bethard et al. [8], pp. 401–407.

[39] Peter Krejzl and Josef Steinberger (2016), UWB at SemEval-2016 task 6: Stance detection. In: Bethard et al. [8], pp. 408–412.

[40] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D. Blondel, and Jari Saramäki (2012), Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):4.

[41] Mirko Lai, Cristina Bosco, Viviana Patti, and Daniela Virone (2015), Debate on political reforms in Twitter: A hashtag-driven analysis of political polarization. In: *Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics*, 19-21 October 2015, Paris, France (DSAA 2015), pp. 1–9. IEEE, Piscataway, NJ, USA.

[42] Mirko Lai, Alessandra Teresa Cignarella, and Delia Irazú Hernández Farías (2017), iTACOS at IberEval2017: Detecting stance in Catalan and Spanish tweets. In: Martínez et al. [54], pp. 185–192.

[43] Mirko Lai, Delia Irazú Hernández Farías, Viviana Patti, and Paolo Rosso (2016), Friends and enemies of Clinton and Trump: Using context for detecting stance in political tweets. In: *Proceedings of the 15th Mexican International Conference on Artificial Intelligence*, 23–28 October 2016, Cancún, Mexico (MICAI 2016), pp. 155–168. Springer International Publishing, Cham, Germany.

[44] Mirko Lai, Rosa Meo, Rossano Schifanella, and Emilio Sulis (2018), The role of the network of matches on predicting success in table tennis. *Journal of Sports Sciences*, 0(0):1–8.

[45] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso (2018), Stance evolution and Twitter interactions in an italian political debate. In: Max Silberztein, Faten Atigui, Elena Kornyshova, Elisabeth Métais, and Farid Meziane, *Natural Language Processing and Information Systems*, 13-15 June 2018, Paris, France (NLDB 2018), pp. 15–27. Springer International Publishing, Cham, Switzerland.

[46] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso (2017), Extracting graph topological information and users' opinion. In: *Proceeding of the 8th International Conference of the CLEF Association*, 11–14 September 2017, Dublin, Ireland (CLEF 2017), pp. 112–118. Springer International Publishing, Cham, Germany.

[47] Paul F. Lazarsfeld and Robert K. Merton. Friendship as a social process: A substantive and methodological analysis. In: *Freedom and Control in Modern Society*, pp. 18–66. Van Nostrand, New York, 1954.

[48] David Lazer, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne (2009), Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723.

[49] Maria Liakata, Jee-Hyub Kim, Shyamasree Saha, Janna Hastings, and Dietrich Rebholz-Schuhmann (2012), Three hybrid classifiers for the detection of emotions in suicide notes. *Biomedical informatics insights*, 5(Suppl. 1):175.

[50] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann (2006), Which side are you on?: Identifying perspectives at the document and sentence levels. In: *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 08-09 June 2006, New York City, New York (CoNLL-X 2006), pp. 109–116. Association for Computational Linguistics, Stroudsburg, PA, USA.

[51] Bing Liu (2012), Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

[52] Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler (2016), IUCL at SemEval-2016 Task 6: An ensemble model for stance detection in Twitter. In: Bethard et al. [8], pp. 394–400.

[53] Miguel Maldonado and Vicenta Sierra (2016), Twitter predicting the 2012 US Presidential Election?: Lessons learned from an unconscious value co-creation platform. *Journal of Organizational and End User Computing*, 28(3):10–30.

[54] Raquel Martínez, Gonzalo Julio, Paolo Rosso, Soto Montalvo, and Jorge Carrillo-de Albornoz (2017). *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017) co-located with 33th Conference of the Spanish Society for Natural Language Processing (SEPLN 2017)*, volume 1881 of *CEUR Workshop Proceedings*, 19 September 2017, Murcia, Spain (CEUR Workshop Proceedings 2017), Aachen, Germany. CEUR-WS.org.

[55] Miller McPherson, Lynn Smith-Lovin, and James M Cook (2001), Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

[56] Enza Messina, Elisabetta Fersini, and Joe Zammit-Lucia. All Atwitter about Brexit: Lessons for the Election campaigns, 2017.

[57] Robert M. Millar (2005), *Language, Nation and Power: An Introduction*. Basingstoke: Palgrave Macmillan, London, United Kingdom.

[58] Amita Misra, Brian Ecker, Theodore Handleman, Nicolas Hahn, and Marilyn Walker (2016), NLDS-UCSC at SemEval-2016 task 6: A semi-supervised approach to detecting stance in tweets. In: Bethard et al. [8], pp. 420–427.

[59] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry (2016), A dataset for detecting stance in tweets. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, May 2016, Portorož, Slovenia (LREC 2016). European Language Resources Association, Paris, France.

[60] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry (2016), SemEval-2016 task 6: Detecting stance in tweets. In: Bethard et al. [8], pp. 31–41.

[61] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu (2013), NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation*, 14-15 June 2013, Atlanta, GA, USA (SemEval 2013), pp. 321–327. Association for Computational Linguistics, Stroudsburg, PA, USA.

[62] Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko (2017), Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

[63] Saif M Mohammad and Peter D. Turney (2013), Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

[64] Saif M. Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin (2015), Sentiment, emotion, purpose, and style in electoral tweets. *Inf. Process. Manage.*, 51(4):480–499.

[65] Mark Newman (2010), *Networks: An Introduction*. Oxford university press, Oxford, UK.

[66] Mark E.J. Newman and Michelle Girvan (2004), Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113.

[67] Finn Årup Nielsen (2011), A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, 30 May 2011, Heraklion, Crete, Greece (CEUR Workshop Proceedings 2011), pp. 93–98. CEUR-WS.org, Aachen, Germany.

[68] M. Nissim and V. Patti. Chapter 3 - semantic aspects in sentiment analysis. In: Federico Alberto Pozzi, Elisabetta Fersini, Enza Messina, and Bing Liu, *Sentiment Analysis in Social Networks*, pp. 31 – 48. Morgan Kaufmann, Boston, 2017.

[69] Yoann Pitarch Romaric Besançon Ophélie Fraisier, Guillaume Cabanac and Boughanem Mohand (2018), Stance classification through proximity-based

community detection. In: *Proceedings of the 29th ACM Conference on Hypertext and Social Media*, 9-12 July 2018, Baltimore, Maryland, USA (ACM Hypertext 2018). Association for Computing Machinery, New York, NY, USA.

[70] Bo Pang and Lillian Lee (2008), Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135.

[71] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay (2016), JU_NLP at SemEval-2016 Task 6: Detecting Stance in Tweets using Support Vector Machines. In: Bethard et al. [8], pp. 440–444.

[72] James W. Pennebaker, Martha E. Francis, and Roger J. Booth (2001), *Linguistic inquiry and word count: LIWC 2001*, volume 71. Lawrence Erlbaum Associates, Mahwah,NJ,USA.

[73] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit (2016), SemEval-2016 task 5: Aspect based sentiment analysis. In: Bethard et al. [8], pp. 19–30.

[74] Francisco Rangel, Paolo Rosso, Martin Potthast, and Benno Stein (2017), Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in Twitter. In: *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum. Vol-1866*, 11-14 September 2017, Dublin, Ireland (CLEF 2017). CEUR-WS.org, Aachen, Germany.

[75] Francisco Manuel Rangel Pardo and Paolo Rosso (2013), Use of language and author profiling: Identification of gender and age. *Natural Language Processing and Cognitive Science*, 177.

[76] Francisco Manuel Rangel Pardo, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches (2013), Overview of the author profiling task at PAN 2013. In: *Working Notes for CLEF 2013 Conference. Vol-1179*, 23-26 September 2013, Valencia, Spain. CEUR-WS.org, Aachen, Germany.

[77] Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans (2015), Overview of the 3rd author profiling task at PAN 2015. In: *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. Vol-1391*, 8-11 September 2015, Toulouse, France. CEUR-WS.org, Aachen, Germany.

[78] Francisco Manuel Rangel Pardo, Paolo Rosso, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daeleman, et al. (2014), Overview of the 2nd author profiling task at PAN 2014. In: *Working Notes for CLEF 2014 Conference*, volume 1180, 15-18 September 2014, Sheffield, UK. pp. 898–927. CEUR-WS.org, Aachen, Germany.

[79] Francisco Manuel Rangel Pardo, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein (2016), Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum*, volume 1609, 5-8 September 2016, Évora, Portugal. pp. 750–784. CEUR-WS.org, Aachen, Germany.

[80] Erzsébet Ravasz, Anna Lisa Somera, Dale A. Mongru, Zoltán N. Oltvai, and Albert-László Barabási (2002), Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

[81] Helmut Schmid (1994), Part-of-speech tagging with neural networks. In: *Proceedings of the 15th conference on Computational linguistics*, 05 - 09 August 1994, Kyoto, Japan (COLING 1994), pp. 172–176. Association for Computational Linguistics, Association for Computational Linguistics, Stroudsburg, PA, USA.

[82] Helmut Schmid (1995), Treetagger – a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

[83] Elliot Schumacher and Maxine Eskenazi (2016), A readability analysis of campaign speeches from the 2016 US Presidential campaign. *CoRR*, abs/1603.05739.

[84] Rob Shields (1741), Solutio problemat is ad geometriam situs pertinentis. commentarii. *Academiae Scientiarum Imperialis Petropolitanae*, 8(4-5):128–140.

[85] Parinaz Sobhani, Saif Mohammad, and Svetlana Kiritchenko (2016), Detecting stance in tweets and analyzing its interaction with sentiment. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, 11-12 August 2016, Berlin, Germany (*SEM 2016), pp. 159–169. Association for Computational Linguistics, Stroudsburg, PA, USA.

[86] Swapna Somasundaran and Janyce Wiebe (2009), Recognizing stances in online debates. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 02 - 07 August 2009, Suntec, Singapore (AFNLP 2009), pp. 226–234. Association for Computational Linguistics, Stroudsburg, PA, USA.

[87] Marco Stranisci, Cristina Bosco, Delia Irazú Hernández Farías, and Viviana Patti (2016), Annotating sentiment and irony in the online Italian political debate on #labuonascuola. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, 23-28 May 2016, Portorož, Slovenia (LREC 2016), pp. 2892–2899. European Language Resources Association, Paris, France.

[88] Emilio Sulis, Delia Irazú Hernández Farías, Paolo Rosso, Viviana Patti, and Giancarlo Ruffo (2016), Figurative messages and affect in Twitter: Differences between #irony, #sarcasm and #not. *Knowledge-Based Systems*, 108:132 – 143.

[89] Cass R. Sunstein (2002), The law of group polarization. *Journal of political philosophy*, 10(2):175–195.

[90] James Joseph Sylvester (1878), On an application of the new atomic theory to the graphical representation of the invariants and covariants of binary quantics, with three appendices. *American Journal of Mathematics*, 1(1):64–104.

[91] Mariona Taulé, M. Antònia Martí, Francisco Manuel Rangel Pardo, Paolo Rosso, Cristina Bosco, and Viviana Patti (2017), Overview of the task of stance and gender detection in tweets on catalan independence at ibereval 2017. In: Martínez et al. [54], pp. 157–177.

[92] Yannis Theocharis and Will Lowe (2016), Does Facebook increase political participation? Evidence from a field experiment. *Information, Communication & Society*, 19(10):1465–1486.

[93] Martin Tutek, Ivan Sekulic, Paula Gombar, Ivan Paljak, Filip Culinovic, Filip Boltuzic, Mladen Karan, Domagoj Alagić, and Jan Šnajder (2016), TakeLab at SemEval-2016 Task 6: Stance Classification in Tweets Using a Genetic Algorithm Based Ensemble. In: Bethard et al. [8], pp. 464–468.

[94] Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy (2016), DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs. In: Bethard et al. [8], pp. 413–419.

[95] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang (2016), pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection. In: Bethard et al. [8], pp. 384–388.

[96] Lilian Weng, Márton Karsai, Nicola Perra, Filippo Menczer, and Alessandro Flammini (2018), Attention on weak ties in social and communication networks. *Computational Social Science*, pp. 213–228.

[97] Robert West, Hristo S. Paskov, Jure Leskovec, and Christopher Potts (2014), Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2:297–310.

[98] Cynthia Whissell (2009), Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports*, 105(2):509–521.

[99] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005), Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, Vancouver, British Columbia, Canada (HLT 2005), pp. 347–354. Association for Computational Linguistics, Stroudsburg, PA, USA.

[100] Michael Wojatzki and Torsten Zesch (2016), tl.uni-due at SemEval-2016 task 6: Stance detection in social media using stacked classifiers. In: Bethard et al. [8], pp. 428–433.

[101] Kaiquan Xu, Jiexun Li, and Stephen Shaoyi Liao (2011), Sentiment community detection in social networks. In: *Proceedings of the 2011 iConference*, 2011, Seattle, WA, USA (iConference 2011), pp. 804–805. Association for Computing Machinery, New York, NY, USA.

[102] Wayne W. Zachary (1977), An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473.

[103] Guido Zarrella and Amy Marsh (2016), MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In: Bethard et al. [8], pp. 458–463.

[104] Zhihua Zhang and Man Lan (2016), ECNU at SemEval 2016 Task 6: Relevant or Not? Supportive or Not? A Two-step Learning System for Automatic Detecting Stance in Tweets. In: Bethard et al. [8], pp. 451–457.