

## 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 textuais i contextuais 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.