Gwerthusiad o Ddiogelwch Ar-lein: trafodaeth ar ddosbarthiad iaith casineb a mesurau diogelwch

Cyfres Papurau Trafod Economaidd

Wedi'i chyhoeddi 12 Mawrth 2024

Trosolwg

Fel rheoleiddiwr diogelwch ar-lein a llwyfannau rhannu fideos, mae angen tystiolaeth arnom i ddeall pa mor effeithiol yw mesurau diogelwch o ran lleihau profiadau pobl o gynnwys niweidiol ar-lein.[1]

Mae diogelwch ar-lein yn gyfundrefn newydd mewn diwydiant deinamig lle bydd gwasanaethau, niwed a mesurau diogelwch newydd yn dod i'r amlwg. Bydd angen i ni ddiweddaru ein dealltwriaeth o niwed, defnyddwyr, gwasanaethau sy'n cael eu rheoleiddio a mesurau diogelwch yn barhaus, a diweddaru ein dull gweithredu ar gyfer polisi ar sail y dystiolaeth hon.

Yn y papur hwn, rydym yn trafod y llenyddiaeth sy'n bodoli ar ba mor effeithiol yw amrywiaeth o fesurau diogelwch a ddefnyddir gan rai llwyfannau i leihau casineb ar-lein. Rydym yn crynhoi'r canfyddiadau allweddol, yn tynnu sylw at fylchau, ac yn cynnig awgrymiadau i lywio cyfeiriad ymchwil yn y dyfodol. Yn benodol, rydym yn tynnu sylw at bwysigrwydd asesu pa mor gywir yw dulliau dosbarthu iaith casineb drwy gynnal ein dadansoddiad ein hunain o gywirdeb dulliau dosbarthu iaith casineb a ddefnyddir yn aml, ac archwilio'r goblygiadau ar gyfer ymchwil ar effeithiolrwydd mesur diogelwch.

Drwy rannu'r papur hwn, ein nod yw ysgogi trafodaethau ar ddatblygu methodolegau cadarn ar gyfer gwerthuso mesurau diogelwch. Rydym hefyd yn ceisio codi ymwybyddiaeth o'r heriau sy'n gysylltiedig â mynd i'r afael ag iaith casineb ar-lein, gan feithrin trafodaeth gyhoeddus fwy gwybodus a manwl ynglŷn ag effaith mesurau diogelwch sy'n delio ag iaith casineb, a sut gellir mesur yr effeithiau hyn.

Beth rydym wedi'i ganfod – yn gryno

Mae tystiolaeth yn y llenyddiaeth academaidd yn awgrymu y gall mesurau diogelwch leihau iaith casineb ar-lein, er bod yr effeithiau'n amrywio'n fawr dros amser a bod y dystiolaeth yn anghyflawn

Mae arbrofion maes a methodolegau lled-arbrofol wedi dangos gostyngiadau sylweddol mewn iaith casineb yn sgil amrywiaeth o fesurau diogelwch, ee ysgogiadau, gwahardd cymunedau neu unigolion, neu ddefnyddio system adrodd y llwyfan i ddileu cynnwys cas. Mae maint yr effeithiau'n amrywio'n fawr ac mae'n ymddangos eu bod yn dibynnu ar nodweddion pob llwyfan.

Mae nifer o fylchau yn y llenyddiaeth o hyd. Mae'r llenyddiaeth wedi canolbwyntio ar set gul o lwyfannau a mesurau diogelwch. Mae llai o ddealltwriaeth o sut byddai mesurau diogelwch ar un llwyfan yn effeithio ar ymddygiad defnyddwyr ar lwyfan arall. Nid yw perfformiad technegau dosbarthu iaith casineb 'oddi ar y silff' yn cael ei asesu'n ddigonol, a gallai hyn wneud canfyddiadau'r ymchwil yn llai cadarn.

Mae'n hollbwysig deall perfformiad dulliau dosbarthu iaith casineb awtomatig wrth ganfod iaith casineb

[1] Rydym yn disgrifio unrhyw ymyriad sy'n cael ei wneud gan lwyfannau ar-lein i leihau profiad defnyddwyr o niwed ar-lein fel mesur diogelwch, er enghraifft ysgogiad i annog defnyddwyr i bostio sylwadau mwy parchus neu wahardd unigolyn o'r llwyfan.

Dulliau dosbarthu iaith casineb yw ffyrdd awtomatig o nodi a yw testun yn cael ei ystyried yn iaith casineb ai peidio. Nid oedd y rhan fwyaf o'r ymchwil a adolygwyd gennym ac a oedd yn defnyddio'r dulliau dosbarthu hyn yn asesu a oedd dulliau dosbarthu iaith casineb yn nodi iaith casineb yn gywir.

Mae llawer o astudiaethau'n defnyddio dulliau dosbarthu iaith casineb generig sydd wedi'u hyfforddi ar ystod eang o setiau data, a elwir yn ddulliau dosbarthu oddi ar y silff. Gall y data hyfforddi (sef y data a ddefnyddir i addysgu'r meddalwedd pa iaith y dylid ei dosbarthu fel iaith casineb) ar gyfer y dulliau dosbarthu hyn fod yn wahanol i'r data a ddefnyddir yn yr astudiaeth werthuso, a gallai hyn arwain at ddosbarthu iaith casineb yn anghywir. Gallai dosbarthiad iaith casineb amhriodol arwain wedyn at ganlyniadau camarweiniol ar effeithiolrwydd mesur diogelwch.

I fynd i'r afael â hyn, byddai angen i ymchwilwyr asesu perfformiad pob un o'r dulliau dosbarthu a ddefnyddir mewn perthynas â'r data iaith ar y llwyfan o ddiddordeb i gadarnhau a oedd y dulliau dosbarthu iaith casineb oddi ar y silff wedi nodi iaith casineb yn gywir ai peidio. Byddai hyn yn rhoi mwy o hyder mai deillio o'r mesur diogelwch y mae newidiadau sylweddol yn y duedd o iaith casineb, yn hytrach na bod yn duedd ar hap sy'n digwydd yn sgil dosbarthu iaith casineb yn anghywir; neu nad yw absenoldeb newid yn digwydd oherwydd y dewis o ddull dosbarthu, y dewis o iaith a ddefnyddir i hyfforddi'r dull dosbarthu hwnnw, neu'r rhagfarn sy'n gynhenid yn y dull dosbarthu.

Fe wnaethom gynnal ein hasesiad ein hunain o ba mor gywir oedd dau ddull dosbarthu a ddefnyddir yn aml i nodi iaith casineb, ac archwiliwyd a oedd y camgymeriadau a gafodd eu gwneud gan ddulliau dosbarthu yn rhai a gafodd eu gwneud ar hap neu ynteu'n rhai a gafodd eu gwneud yn systematig.

Fe wnaethom ddefnyddio'r dulliau dosbarthu iaith casineb gyda'i set ddata'r prawf HateXplain – sef set ddata o sylwadau ar y cyfryngau cymdeithasol gyda labeli ar gyfer iaith casineb a labeli cyffredinol, gan nodi targed yr iaith casineb. Fe wnaethom asesu perfformiad dau ddull dosbarthu iaith casineb a ddefnyddir yn aml: Rhyngwyneb Rhaglenni Cymwysiadau Google Perspective – y dull dosbarthu 'oddi ar y silff' a ddefnyddir amlaf a'r model HateXplain, a gafodd ei hyfforddi ar ddata tebyg i set ddata'r prawf HateXplain. Nod yr astudiaeth achos hon oedd cymharu perfformiad dull dosbarthu 'oddi ar y silff' â dull dosbarthu a gynlluniwyd yn benodol ar gyfer y set ddata.[2]

Daeth i'r amlwg i ni mai'r dull dosbarthu a ddyluniwyd yn benodol ar gyfer y set ddata, HateXplain, oedd yn perfformio orau – gan nodi'n union y rhan fwyaf o'r iaith casineb yn y set ddata, tra bod y dull dosbarthu 'oddi ar y silff' wedi methu â nodi'r rhan fwyaf o'r sylwadau iaith casineb yn y set ddata. Gwelsom hefyd fod perfformiad y dulliau dosbarthu yn amrywio yn dibynnu ar darged yr iaith casineb. Roedd y dulliau dosbarthu yn gwneud mwy o gamgymeriadau wrth nodi iaith casineb wedi'i thargedu at rai grwpiau ethnig o'i gymharu ag eraill.

Rydym yn tynnu sylw at y ffaith, os nad yw camgymeriadau dulliau dosbarthu iaith casineb 'oddi ar y silff' yn cael eu profi ac nad ydynt yn cael eu deall, yna gall eu defnyddio arwain at ragfarn wrth ddadansoddi casgliadau achosol, gan ei gwneud yn anoddach gwerthuso a yw ymyriad wedi cael yr effaith a ddymunir. Yn ein trafodaeth, rydym yn cyfeirio at ddulliau y gellir eu defnyddio i astudio a chywiro'r rhagfarn hon.

[2] I gael manylion am y set ddata a'r model HateXplain, gweler Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P. a Mukherjee, A., 2021, Mai. Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence (Cyf. 35, Rhif 17, tt. 14867-14875); ac, i gael manylion am ddull dosbarthu Google Perspective, gweler https://perspectiveapi.com/