ISD | Powering solutions to extremism and polarisation

Casineb y Genedl

Mapio Tirwedd yr Iaith y Gellir Arsylwi ei bod yn Gredadwy o Atgas ar y Cyfryngau Cymdeithasol

Jacob Davey, Carl Miller, Jakob Guhl

RHYBUDD: Mae'r adroddiad hwn yn cynnwys disgrifiadau o bostiadau a allai beri gofid i ddarllenwyr. Mae'r rhain yn cynnwys samplau o iaith casineb sy'n cynnwys iaith atgas a dad-ddyneiddiol sy'n targedu menywod a chymunedau lleiafrifol.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | Llundain | Paris | Washington DC

Cynnwys

Trosolwg 4

Crynodeb Gweithredol 5

Trosolwg

Mae'r adroddiad hwn yn darparu trosolwg o'r negeseuon iaith Saesneg cyhoeddus a gasglwyd o Facebook, Instagram, Twitter, Reddit a 4chan yn ystod mis Awst 2022 yr ydym yn eu hystyried yn 'gredadwy o atgas'. Ystyr hyn yw y tybir mai o leiaf un o'r dehongliadau rhesymol o'r neges yw ei bod yn ceisio dad-ddyneiddio, pardduo, eithrio, aflonyddu neu fygwth unigolyn neu gymuned ar sail nodwedd warchodedig, mynegi dirmyg neu ffieidd-dod tuag ato neu annog trais yn ei erbyn. Deellir mai nodweddion gwarchodedig yw hil, tarddiad cenedlaethol, anabledd, ymlyniad crefyddol, cyfeiriadedd rhywiol, rhyw, neu hunaniaeth o ran rhywedd.

Crynodeb Gweithredol

Mae ymchwilio i gasineb ar gyfryngau cymdeithasol yn un o'r mathau pwysicaf ond anoddaf o ymchwil i'w gwneud ar-lein. Mae'n ffenomen y mae'n hollbwysig ei deall er mwyn disgrifio'n llawn natur gofodau ar-lein a phrofiadau gwahanol gymunedau sy'n byw ynddynt. Ar y llaw arall - fel y mae'r adroddiad hwn yn ei drafod - mae adnabod iaith casineb yn gywir ar draws ystod o lwyfannau mewn ffordd synhwyrol, onest a chadarn yn her ymchwil aruthrol, a hynny o safbwynt diffiniadol a thechnolegol.

Penllanw yw'r adroddiad hwn o brosiect ymchwil oedd â'r nod o adnabod iaith casineb ar Facebook, Instagram, Reddit, Twitter a bwrdd /pol/ 4chan dros gyfnod o fis, a hefyd nodi'r ystyriaethau amrywiol o ran data, methodoleg a dull epistemig sy'n gysylltiedig â'r arfer ymchwil hwn. Mae gormod o lawer o weithgarwch ar draws y cyfryngau cymdeithasol i fedru byth cynnal dadansoddiad dynol cynhwysfawr, ac wrth wraidd yr ymdrech ymchwil hon oedd hyfforddi, defnyddio a gwerthuso offer prosesu iaith naturiol (NLP) i ganfod iaith casineb ar sail algorithm. Mae dosbarthu iaith casineb yn awtomataidd wedi bod yn destun diddordeb academaidd a masnachol ers nifer o flynyddoedd bellach, ac er mwyn adeiladu ar y cynnydd a wnaed gan grwpiau eraill, fe wnaethom gyfuno llawer o fodelau dosbarthu casineb gyda'i gilydd yn fodel o fodelau, neu 'ensemble'. Mae cryfderau a chyfyngiadau'r dull hwn hefyd yn cael eu trafod isod, ac mae'n hanfodol bod y canfyddiadau a gyflwynir yn yr adroddiad hwn yn cael eu darllen gan gadw'r cafeatau hyn mewn cof.

Er bod yr ymchwil yn cwmpasu nifer o lwyfannau, maent yn wahanol iawn i'w gilydd, ac ni ddylid gwneud cymariaethau rhwng llwyfannau ar sail yr astudiaeth hon**.** Mae Twitter, 4Chan, Instagram, Facebook a Reddit yn amrywio o ran eu maint, pwy sy'n eu defnyddio a sut maen nhw'n ffitio i fywydau pobl. Maent yn wahanol hefyd o ran y polisïau sydd ganddynt (os o gwbl) ar iaith casineb, beth yw'r polisïau hynny, a sut maent yn cael eu gorfodi. Yn bwysicaf oll efallai, dylanwadir yn fawr ar raddau'r casineb a nodwyd yn ein hymchwil ar gyfer pob llwyfan gan y graddfeydd data y mae pob llwyfan yn eu darparu i'w darganfod a'u casglu, y rhyngweithio amrywiol rhwng ein meini prawf cywain data gyda phob llwyfan, perfformiad adalw amrywiol ein system ar bob llwyfan, a gweithgarwch pob llwyfan o ran dileu cynnwys, yn unol â chanllawiau'r llwyfan benodol honno.

Mae iaith casineb yn dibynnu'n fawr ar gyd-destun. Yn aml, nid yw'n bosibl hyd yn oed i ddadansoddwyr dynol benderfynu a yw postiad penodol mewn gwirionedd yn atgas o'i ystyried allan o'i gyd-destun. Yn ystod y dadansoddiad ar gyfer yr adroddiad hwn, yn aml nid oedd gan ddadansoddwyr a oedd yn ceisio asesu a oedd postiad yn atgas ai beidio ddigon o wybodaeth hanfodol am hunaniaeth anfonwr a derbynnydd postiad, na'r cyd-destun ehangach o'i greu. Am fod termau atgas yn aml yn cael eu hailddefnyddio a'u hawlio'n ôl gan eu grwpiau targed, mae'n anodd pennu'n hyderus felly beth yw'r bwriad y tu ôl i'r defnydd o sarhad o'r fath. Ar yr un pryd, gall teimladau atgas gael eu cyfathrebu hefyd mewn ffordd fwy amwys, ymhlyg a chynnil.

Oherwydd y cafeatau hyn, mae côdwyr dynol a pheiriannau fel ei gilydd yn cael trafferth gydag achosion ymylol lle mae'n ansicr a yw postiad yn atgas ai beidio. I fynd i'r afael â'r her hon, rydym yn cyflwyno'r categori 'credadwy o atgas' i ddisgrifio postiadau y mae dehongliadau lluosog yn bodoli ar eu cyfer ac un dehongliad rhesymol yw ei fod yn wir yn atgas. Cafodd y postiadau hyn eu codio fel rhai a oedd yn gredadwy o atgas a dyna sut y cyfeirir atynt drwy gydol yr adroddiad llawn.

Mae'r ymchwil yn rhoi cipolwg ar iaith sy'n gredadwy o atgas ar gyfryngau cymdeithasol felly nid yw'n gynrychioliadol, nac yn rhydd o'r cyfyngiadau sy'n gysylltiedig â'r methodolegau y mae'r ymchwil yn eu defnyddio ac yn rhan annatod ohonynt. Fodd bynnag, credwn ei bod yn parhau i fod yn hanfodol, yn bwysig ac yn berthnasol wrth i gymdeithas barhau i gyd-drafod sut i adeiladu amgylcheddau digidol sydd hefyd yn oddefgar ac amrywiol.

Canfyddiadau allweddol

Dros fis Awst 2022, mis ein hastudiaeth:

- Casglwyd 3,140,324 o negeseuon cyhoeddus rhwng 01 Awst 2022 a 31 Awst 2022 a anfonwyd ar 4chan, Facebook, Instagram, Reddit a Twitter a oedd yn cynnwys o leiaf un o 334 o allweddeiriau neu ymadroddion allweddol sy'n gysylltiedig ag iaith casineb a nodwyd gennym.

- O'r rhain, roedd 422,681 o negeseuon wedi'u dosbarthu fel rhai 'credadwy o atgas', hynny yw, lle mai un dehongliad rhesymol o'i ystyr oedd ei fod yn ceisio dad-ddyneiddio, pardduo, eithrio, aflonyddu neu fygwth unigolyn neu gymuned ar sail nodwedd warchodedig, mynegi dirmyg neu ffieidd-dod tuag ato neu annog trais yn ei erbyn. At

ddibenion yr ymchwil hon, diffiniwyd nodweddion gwarchodedig gennym fel hil, tarddiad cenedlaethol, anabledd, ymlyniad crefyddol, cyfeiriadedd rhywiol, rhyw, neu hunaniaeth o ran rhywedd.

Ar draws llwyfannau, fe wnaethom nodi:

- **394,753** o negeseuon credadwy o atgas ar Twitter.
- **26,085** o negeseuon credadwy o atgas ar 4Chan.
- **1,540** o negeseuon credadwy o atgas ar Facebook.
- **162** o negeseuon credadwy o atgas ar Instagram.
- **141** o negeseuon credadwy o atgas ar Reddit.

- **Nid yw'r niferoedd hyn yn gyfystyr â graddau llawn y casineb ar bob llwyfan, nac yn sampl gynrychioliadol o bob llwyfan.** Yn unol â hynny, dylid dehongli ein canfyddiadau i olygu bod o leiaf y nifer yma o negeseuon sy'n gredadwy o atgas yn bresennol ar y llwyfannau hyn yn ystod mis Awst 2022, yn hytrach na'u bod yn arwydd o faint absoliwt o gynnwys atgas.

- **Mae'n hanfodol bod y canfyddiadau hyn yn cael eu hystyried yn erbyn realiti maint y llwyfan a mynediad at ddata.** Mae graddau'r negeseuon atgas ar Twitter yn ganlyniad i'r faith bod y llwyfan honno yn darparu llawer mwy o fynediad at ddata i ddadansoddwyr yn ystod yr ymchwil hon. Oherwydd anghysondebau mewn mynediad at ddata a nifer o resymau eraill, **ni ddylid** defnyddio'r canfyddiadau hyn i wneud cymariaethau am raddau'r iaith casineb ar bob llwyfan.

- **Mae natur iaith casineb yn amrywio gan ddibynnu ar normau diwylliannol llwyfannau.** Awgrymodd dadansoddiad ansoddol o negeseuon credadwy o atgas a samplwyd ar hap fod termau o sarhad yn cael eu defnyddio mewn trafodaeth reolaidd ar 4chan, gan awgrymu bod rhai defnyddwyr ar y llwyfan wedi normaleiddio'r defnydd o iaith atgas. Nododd yr un dadansoddiad fod casineb ar Facebook, Instagram a Twitter i'w weld yn ymddangos yn bennaf yn y defnydd o sarhad, ac mewn iaith fwy amwys, megis cyflwyno damcaniaethau cynllwyn sy'n pardduo cymunedau lleiafrifol.

- **Nid yw iaith casineb yn ysgogi lefelau uwch o ryngweithio na negeseuon nad ydynt yn atgas.** Ar Reddit a Twitter, cyflawnodd negeseuon nad oeddent yn atgas fwy o hoffiadau fesul postiad mewn gwirionedd. Darlun mwy amwys a gafwyd ar y llwyfannau sy'n eiddo i Meta, am fod negeseuon atgas yn derbyn llai o achosion o hoffi ac ymateb ar gyfartaledd na negeseuon nad oeddent yn atgas ar Facebook ond cawsant eu rhannu'n fwy a gwnaed mwy o sylwadau arnynt.

- **Gwelwyd bod rhai o'r negeseuon credadwy o atgas a nodwyd yn parhau ar y llwyfannau fis yn ddiweddarach, tra bod rhai eraill yn anhygyrch.**
  - Ar Reddit nid oedd 26% o negeseuon credadwy o atgas ar gael mwyach fis ar ôl cael eu casglu.
  - Ar Twitter, nid oedd 18.2% o negeseuon credadwy o atgas ar gael mwyach fis ar ôl cael eu casglu.
  - Ar Instagram nid oedd 14.8% o negeseuon credadwy o atgas ar gael mwyach fis ar ôl cael eu casglu.
  - Ar Facebook, nid oedd 11.5% o negeseuon credadwy o atgas ar gael mwyach fis ar ôl cael eu casglu.

Wrth ddarllen y canfyddiadau hyn, mae'n bwysig ystyried nifer o gafeatau. Manylir arnynt isod, ond maent yn cynnwys y canlynol:

- **Mae'r dosbarthwr peirianyddol yn cyflwyno canlyniadau positif ffug i'n canlyniadau.** Nodir hyn trwy werthuso'r model ar gyfer 'manwl gywirdeb'. Mae manwl gywirdeb y model wedi'i fesur fel: 91% o fanwl gywirdeb ar 4chan; 71% o fanwl gywirdeb ar Twitter; a 63% ar Facebook ac Instagram. Unigolyn a ddosbarthodd y data Reddit a gasglwyd. Mae hyn yn golygu y canfuwyd nad oedd oddeutu tair o bob deg o'r negeseuon Twitter, pedair o bob deg o'r negeseuon Facebook ac Instagram, ac un o bob deg o'r negeseuon 4chan y rhagfynegodd y model eu bod yn rhai credadwy o atgas, mewn gwirionedd yn gredadwy o atgas o'u gwerthuso gan unigolion dynol. Mae hyn ynddo'i hun yn dangos heriau adnabod iaith casineb – hyd yn oed gyda methodoleg gymhleth a nifer o fodelau – ac mae'r ffaith hon, ynghyd ag anghysondebau o ran mynediad at ddata fel y trafodir uchod yn heriau nodedig y gallai fod angen i astudiaethau yn y dyfodol sydd yn yr un modd yn ceisio dadansoddi iaith casineb ar raddfa ac ar draws llwyfannau ymbalfalu â nhw.

- **Nid yw casgliadau sy'n seiliedig ar allweddeiriau yn creu setiau data cynrychioliadol, ac ni ellir cyffredinoli nac allosod y canlyniadau yma i roi amcangyfrif o gyfanswm y gweithgarwch credadwy o atgas ar bob llwyfan.** Gan hynny, nid yw'n bosibl i ni fesur lefel adalw gyffredinol y llif gwaith. Yn unol â hynny mae'r lefelau gwahanol o fynediad at ddata a ddarperir gan lwyfannau'n cyflwyno her sylfaenol i unrhyw brosiect ymchwil sy'n ceisio cymharu lefelau iaith casineb ar draws

cyfryngau cymdeithasol.

- **Mae gan lwyfannau gwahanol bolisïau cymedroli gwahanol, ac nid yw'r diffiniadau o iaith casineb a ddefnyddir yn yr astudiaeth hon o reidrwydd yn**
- **cynrychioli iaith casineb fel y'i diffinnir gan Delerau Gwasanaeth neu ganllawiau cymunedol unrhyw lwyfan benodol.** Felly, nid yw'r papur yn honni bod bodolaeth barhaus neges gredadwy o atgas o reidrwydd yn gyfystyr â methiant gorfodi'r llwyfan dan sylw.

ISD | Powering solutions to extremism and polarisation

Amman | Berlin | Llundain | Paris | Washington DC