Detekcija SPAM poruka pomoću LSTM-a

Maksim Kos Tomislav Matić Lana Tuković

Hrvoje Ljubas

Jerko Šegvić

Ključne riječi—detekcija, LSTM, neuronske mreže, spam

I. UVOD

Detekcija SPAM poruka je proces identifikacije i filtriranja neželjenih, često masovno poslanih poruka koje se šalju putem različitih komunikacijskih kanala, kao što su e-mailovi, SMS poruke, društvene mreže itd. Cilj detekcije SPAM-a je razdvajanje legitimnih poruka od onih koje korisnici ne žele ili koje predstavljaju potencijalnu prijetnju. Detekcija SPAM poruka u e-mailovima i SMS porukama ključan je aspekt održavanja sigurnosti i učinkovitosti komunikacije putem tih kanala. Neželjene poruke mogu varirati od promotivnih materijala do zlonamjernih prijetnji, stoga je važno razviti učinkovite tehnike za njihovo prepoznavanje.

Klasične karakteristike SPAM poruka mogu uključivati:

- Nepoželjna promocija: ponude proizvoda, usluga ili marketinške kampanje koje niste zatražili.
- Prijevare: pokušaji prevarantskih aktivnosti, poput phishinga, u kojima se pokušava prevariti korisnike kako bi otkrili osobne ili financijske informacije.
- Neprimjeran sadržaj: poruke koje sadrže uvredljiv, neprimjeren ili zlonamjeran sadržaj.
- Lažne nagrade: ponude koje obećavaju nagrade ili povlastice kako bi privukle korisnike.

Detekcija SPAM-a obično uključuje upotrebu različitih tehnika, uključujući:

- Analizu značajki: Praćenje specifičnih obrazaca ili karakteristika koje su često prisutne u SPAM porukama.
- Strojno učenje: Korištenje algoritama strojnog učenja za treniranje modele prepoznavanja SPAM-a na temelju označenih podataka.
- Bijele i crne liste: Održavanje popisa poznatih sigurnih i SPAM adresa kako bi se brže identificirale neželjene poruke.
- Analiza ponašanja: Praćenje ponašanja korisnika ili obrazaca slanja poruka kako bi se identificirale nepravilnosti.
- Aktivno ažuriranje: Redovito ažuriranje sustava detekcije kako bi se prilagodile nove tehnike SPAM-a.

Cilj je poboljšati korisničko iskustvo, povećati sigurnost komunikacije i spriječiti zloupotrebu putem neželjenih poruka. Detekcija SPAM-a često predstavlja neprekidan proces prilagodbe i evolucije kako bi se održala učinkovitost u suočavanju s novim oblicima i taktikama SPAM-a koji se stalno mijenjaju.

II. PREGLED POSTOJEĆIH PRISTUPA

Sustavi za detekciju spam poruka razvijaju se već nekoliko desetljeća, a u nastavku su neki od postojećih:

- detekcija bazirana na pravilima (Rule-based):

Tradicionalni pristupi koriste pravila za identifikaciju SPAM poruka. Ova pravila mogu uključivati analizu ključnih riječi, prisustvo određenih znakova (npr. velika slova, znakovi interpunkcije), ili analizu strukture poruke.

- Bayesian Filtering:

Bayesovski filtri koriste statistički model temeljen na Bayesovom teoremu za klasifikaciju poruka kao SPAM ili ne-SPAM. Ti modeli se treniraju na temelju označenih skupova podataka.

- Machine Learning pristupi:

Klasični strojno učeni modeli, poput Support Vector Machine (SVM) ili Decision Trees, često se koriste za klasifikaciju SPAM-a. Ovi pristupi zahtijevaju ekstrakciju značajki iz poruka.

- Deep Learning pristupi:

Duboko učenje donosi naprednije pristupe, uključujući korištenje povratnih neuronskih mreža (RNN) s LSTM mehanizmom. LSTM omogućuje modelima da bolje prate kontekst i održavaju dugoročne ovisnosti u tekstu.

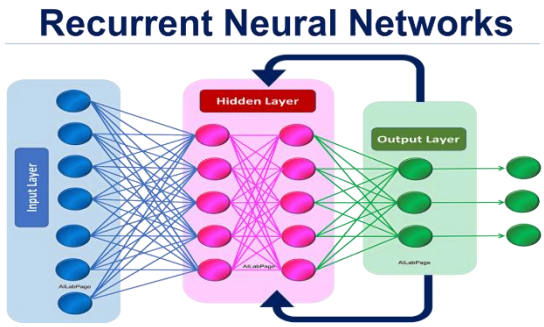
- Ensemble metode:

Kombiniranje više modela, poznato kao metoda ansambla, može poboljšati ukupnu točnost. Na primjer, kombinacija modela baziranog na pravilima i dubokog modela.

III. OPIS RJEŠENJA PROBLEMA

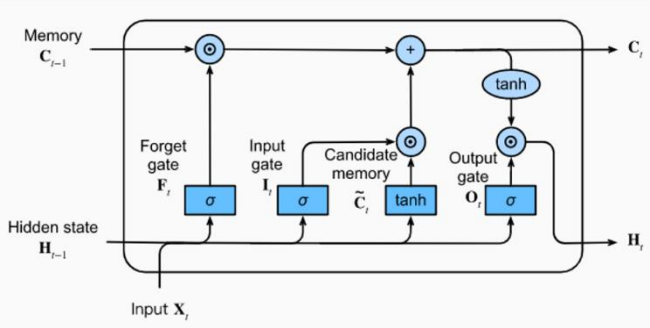
Detekciju SPAM poruka u e-mailovima i SMS porukama izveli smo korištenjem LSTM-a (Long Short-Term Memory). Detekcija SPAM poruka pomoću LSTM-a temelji se na korištenju povratnih neuronskih mreža (RNN). LSTM je poseban tip povratne neuronske mreže dizajniran za obradu i modeliranje sekvencijalnih podataka, kao je tekst.

Povratne neuronske mreže su vrsta neuronskih mreža koje imaju povratne veze. Klasične neuronske mreže nisu pogodne za obradu sekvencijalnih podataka jer ne zadržavaju informacije o prethodnim koracima. RNN pokušavaju riješiti taj problem dodajući povratnu petlju, što im omogućuje održavanje unutarnjeg stanja ili "memorije" koje se može ažurirati s prethodnim koracima sekvence.



https://miro.medium.com/v2/resize:fit:1400/1*5wYNhHV-F7U1hV5pscYz5Q.png

LSTM je vrsta povratne neuronske mreže koja rješava problem dugoročne ovisnosti koji se često pojavljuje u običnim RNN-ovima. Standardne RNN-ove karakterizira problem nestajućeg i eksplodirajućeg gradijenta, što ograničava njihovu sposobnost održavanja dugoročnih zavisnosti. LSTM rješava taj problem uvođenjem posebnih mehanizama poput ćelija s vratima (gates).



https://miro.medium.com/v2/resize:fit:984/1*Mb_L_slY9rjMr8-IADHvwg.png

Prvo je potrebno prikupiti označne podatke. Za e-mailove nam treba skup primjera s oznakom je li taj e-mail SPAM ili nije. Isto tako za SMS poruke nam treba skup primjera SMS poruka s pripadajućom oznakom. Postoje već gotovi skupovi podataka, za e-mailove smo koristili: https://www.kaggle.com/datasets/mfaisalqureshi/spam-email, dok smo za SMS poruke koristili: https://www.kaggle.com/code/basilb2s/sms-spam-detection-using-lstm/input.

Kako bismo mogli obrađivati tekst, potrebno ga je razbiti na manje jedinice koje nazivamo tokeni. Tokeni su osnovne građevne jedinice kojima se tekst razdvaja, a mogu predstavljati riječi, znakove interpunkcije, brojeve ili druge segmente teksta. Tokeniziranje teksta ima ključnu ulogu u pretvaranju neprekinutog niza znakova u strukturu podataka koja je pogodna za analizu i obradu u okviru NLP sustava. Ovaj proces omogućava računalu da bolje razumije i interpretira tekst na temelju značajki svakog pojedinog tokena. Svaki token možemo indeksirati te onda tokenizirani tekst možemo pretvoriti u niz indeksa pripadajućeg tokena. Kraće nizove indeksa je potrebno nadopuniti tako da svi nizovi budu jednake duljine, odnosno sada imamo vektore brojeva istih dimenzija koje možemo dovesti na ulaz neuronske mreže.

Tako obrađeni podaci su zatim izmiješani i raspodijeljeni u skup za treniranje, testiranje i provjeru.

Neuronska mreža započinje sa slojem koji ulazne podatke preslikava u prostor značajki. Slični ulazi će biti preslikani u vektore u prostoru značajki koji se nalaze relativno blizu, dok će vektori različitih ulaza u prostoru značajki biti udaljeniji. Zatim slijedi LSTM sloj, koji sadrži memoriju i stanje te je prethodno objašnjen. Nakon LSTM sloja slijedi „Dropout“ sloj koji služi za regularizaciju te funkcionira na način da nasumično isključuje aktivaciju određenog neurona u sloju, kako se mreža ne bi prenaučila na podacima za treniranje. Taj sloj prisiljava mrežu da se osloni na različite puteve i značajke čime se poboljšava generalizacija na nepoznatim podatcima. Bitno je napomenuti da se taj sloj koristi samo pri treniranju, ne i tijekom evaluacije. Iza tog sloja slijedi potpuno povezani sloj, što znači da je svaki čvor ovog sloja povezan sa svakim čvorom iz prethodnog i sljedećeg sloja. Na poslijetku slijedi još jedan par „Dropout“ sloja te potpuno povezanog sloja dimenzije 1, koji je ujedno i izlaz neuronske mreže.

Sad kad imamo arhitekturu neuronske mreže, potrebno je podesiti hiperparametre. Hiperparametri koje je potrebno podesiti su broj memorijskih jedinica u LSTM sloju, vjerojatnost isključivanja neurona u Dropout sloju, broj neurona u potpuno povezanom sloju te aktivacijske funkcije za LSTM sloj i potpuno povezani sloj. Za svaki hiperparametar napravit ćemo listu vrijednosti koje može poprimiti te ćemo pronaći najbolju kombinaciju hiperparametara pretraživanjem po rešetci. Drugim riječima iterirat ćemo po svim mogućim kombinacijama hiperparametara, trenirati model s pripadnom kombinacijom hiperparametara na skupu za treniranje, zatim provjeriti točnost na skupu za provjeru te ćemo spremiti model s najvećom točnošću. Taj ćemo model smatrati najboljim te bismo njega koristili za klasifikaciju SPAM poruka.

Ista arhitektura neuronske mreže je korištena i za e-mailove i za SMS poruke, razlika je u tome što su trenirane na različitim podacima. Osim što smo provjerili točnost neuronske mreže na pripadnom skupu podataka za validaciju, provjerili smo kako se ponaša mreža trenirana na e-mailova pri klasifikaciji SPAM-a u SMS porukama te kako mreža trenirana na SMS porukama radi s e-mailovima.

IV. OPIS EKSPERIMENTALNIH REZULTATA

U sklopu projekta istrenirane su dvije duboke neuronske mreže sa LSTM(Long Short-term Memory) arhitekturom za klasifikaciju spam poruka.

Svaka od mreža trenirana je na 5000-6000 podataka, međutim, jedna je trenirana na SMS porukama, a druga na e-mailovima.

Proces započinje podjelom podataka u skup za treniranje, validaciju i testiranje te tokenizacijom riječi.

Optimizacijom hiperparametara mreža postignuta je točnost od 99.48% za klasifikaciju SMS-ova te 95.6% za klasifikaciju e-mailova na skupovima za testiranje.

Kao dodatam eksperimentalni validacijski korak, model treniran nad e-mailovima testiran je nad sms-porukama i obratno.
Rezultat toga je da ni jedan od modela nije radio (točnost je bila manja od 50%). Takav rezultat se može objasniti činjenicom da tekstovi od mailova uvijek počinju zaglavljem. Isto tako, mailovi imaju formalniju strukturu i obično su dulji nego SMS poruke.

USPOREDBA REZULTATA S REZULTATIME PRETHODNE LITERATURE

Klasifikacija SPAM-a u SMS porukama rađena je po uzoru na: https://www.kaggle.com/code/basilb2s/sms-spam-detection-using-lstm/notebook. U literaturi u posljednjoj epohi treniranja, pogreška na skupu za treniranje je 0.0092, preciznost na skupu za treniranje je 0.9974, dok je pogreška na skupu za validaciju 0.159, a preciznost na skupu za validaciju 0.9417. U literaturi nije rađeno pretraživanje hiperparametara po rešetci, već su odmah zadani. U našem projektu najbolja preciznost na skupu za validaciju je 0.9905. U literaturi nema preciznosti na podacima za provjeru. Moguće da je razlika uzrokovana pretraživanjem po rešetci, jer smo tražili kombinaciju hiperparametara koja daje najbolje rezultate na skupu za validaciju te taj model sačuvali, dok su u literaturi hiperparametri fiksirani te se treniranjem nastoji minimizirati pogreška na skupu za treniranje. Osim toga drugačije su podijeljeni podaci. U našem projektu ima 3000 uzoraka za treniranje, 2000 uzoraka za validaciju i 572 uzorka za provjeru, dok je u literaturi 80% podataka za treniranje, a 20% za validaciju. Model za detekciju SPAM-a u e-mailovima je rađen po uzoru na model za SMS poruke, samo treniran na e-mailovima, stoga nema referentan projekt za usporedbu rezultata.

ZAKLJUČAK

U provedbi ovog projekta usmjerenog na detekciju spam poruka, implementirane su dvije duboke neuronske mreže temeljene na LSTM (Long Short-term Memory) arhitekturi. Eksperimenti su pružili značajne uvide u sposobnosti modela, s fokusom na klasifikaciju između SMS poruka i e-mailova.

Rezultati istraživanja potvrđuju visoku točnost obećavajućih 99.48% za klasifikaciju SMS poruka i 95.6% za klasifikaciju e-mailova. Međutim, važno je napomenuti da model nije radio kada je treniran na jednom tipu poruka a testiran na drugom. Ovaj pad učinkovitosti ukazuje na specifičnosti strukture i sadržaja između SMS poruka i e-mailova. Ovi rezultati naglašavaju nedostatak prilagodljivosti modela ovisno o vrsti poruka koja se analizira.

U zaključku, unatoč izazovima u klasifikaciji različitih vrsta poruka, projekt je ostvario značajan napredak u detekciji spam poruka korištenjem dubokog učenja s LSTM arhitekturom. Daljnji rad može usmjeriti pažnju na poboljšanja modela kako bi se riješili specifični problemi povezani s raznolikošću poruka, pridonoseći efikasnijem sustavu zaštite od neželjenih poruka.

LITERATURA

https://towardsdatascience.com/spam-detection-in-emails-de0398ea3b48
https://medium.com/@ottaviocalzone/an-intuitive-explanation-of-lstm-a035eb6ab42c
https://en.wikipedia.org/wiki/Long_short-term_memory
https://en.wikipedia.org/wiki/Recurrent_neural_network
https://en.wikipedia.org/wiki/Anti-spam_techniques
https://www.enjoyalgorithms.com/blog/email-spam-and-non-spam-filtering-using-machine-learning