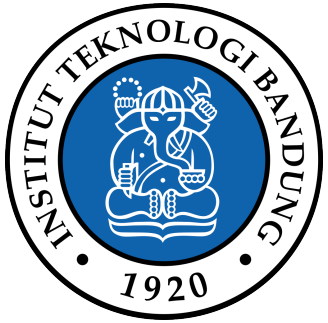


IF6028 Pemrosesan Bahasa Alami

DOTA 2 Toxic Player Detection

Disusun oleh
Arrival Dwi Sentosa 23519035

Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung
2020

LATAR BELAKANG

"Toxic Player"
Tindakan pemain yang merugikan pemain lain:

MASALAH

1. Sengaja bermain jelek
2. Melakukan ujaran tulisan yang tidak baik:
   - Menyalahkan pemain lain
   - Menghina pemain lain
   - Perkataan kasar
   - SARA
   - Sarkasme

DAMPAK

1. Menghambat pemain baru berkembang
2. Pengalaman bermain menjadi buruk:
   - Penurunan jumlah pemain
   - Tujuan refreshing menjadi stress
   - Terbiasa dengan kata kasar

SOLUSI SAAT INI

Valve selaku produsen dari game DOTA 2, Menambahkan fitur laporan untuk melaporkan pemain toxic, akan tetapi:

1. Pemain tidak langsung di hukum
2. Perlu proses review berhari hari

*Karena jumlah pertandingan yang sangat banyak

SOLUSI
DIUSULKAN

DOTA 2 Toxic Player Detection

Bertujuan untuk mendeteksi apakah seorang pemain Dota 2 bersifat toxic atau tidak, berdasarkan rekaman percakapan dalam satu pertandingan.

DATASET

Didapatkan dari percakapan pemain dalam game

Dengan karakteristik sebagai berikut:

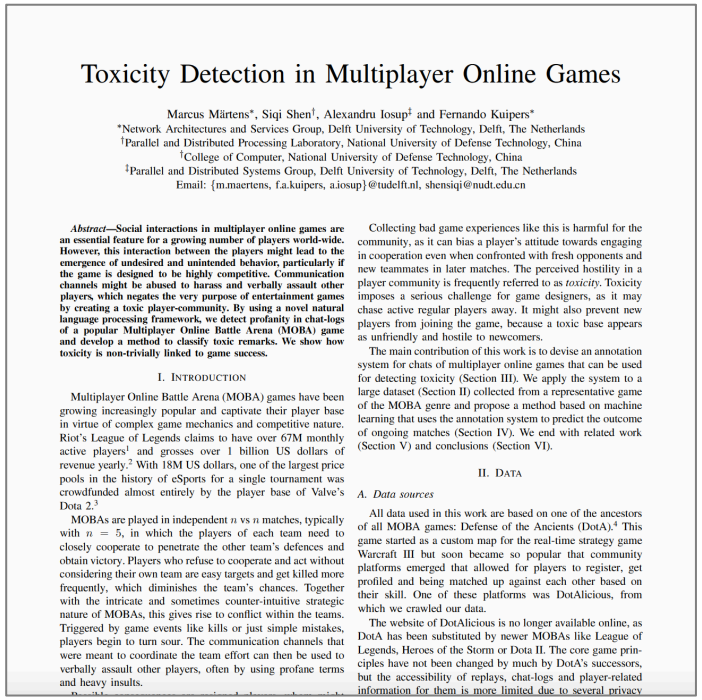
- Mengandung **spelling-error**
- Kebanyakan tidak mengikuti **grammatical structure**
- Mengandung **pidgin language**, Bahasa tidak baku, **Campuran berbagai Bahasa** dalam satu kalimat
- **Istilah atau jargon** yang hanya dikenal dalam DOTA 2
- Pemain menyebut dengan **nama hero pemain lain**

MODUL

- Named Entity Recognition
- Word Embedding
- Text Classification

ALUR APLIKASI

APLIKASI SEJENIS



Toxicity Detection in Multiplayer Online Games

Marcus Märtens*, Siqi Shen†, Alexandru Iosup‡ and Fernando Kuipers*
*Network Architectures and Services Group, Delft University of Technology, Delft, The Netherlands
†Parallel and Distributed Processing Laboratory, National University of Defense Technology, China
†College of Computer, National University of Defense Technology, China
‡Parallel and Distributed Systems Group, Delft University of Technology, Delft, The Netherlands
Email: {m.maertens, f.a.kuipers, a.iosup}@tudelft.nl, shensiqi@nudt.edu.cn

***Abstract*—Social interactions in multiplayer online games are an essential feature for a growing number of players world-wide. However, this interaction between the players might lead to the emergence of undesired and unintended behavior, particularly if the game is designed to be highly competitive. Communication channels might be abused to harass and verbally assault other players, which negates the very purpose of entertainment games by creating a toxic player-community. By using a novel natural language processing framework, we detect profanity in chat-logs of a popular Multiplayer Online Battle Arena (MOBA) game and develop a method to classify toxic remarks. We show how toxicity is non-trivially linked to game success.**

I. Introduction

Multiplayer Online Battle Arena (MOBA) games have been growing increasingly popular and captivate their player base in virtue of complex game mechanics and competitive nature. Riot's League of Legends claims to have over 67M monthly active players[1] and grosses over 1 billion US dollars of revenue yearly.[2] With 18M US dollars, one of the largest price pools in the history of eSports for a single tournament was crowdfunded almost entirely by the player base of Valve's Dota 2.[3]

MOBAs are played in independent n vs n matches, typically with $n = 5$, in which the players of each team need to closely cooperate to penetrate the other team's defences and obtain victory. Players who refuse to cooperate and act without considering their own team are easy targets and get killed more frequently, which diminishes the team's chances. Together with the intricate and sometimes counter-intuitive strategic nature of MOBAs, this gives rise to conflict within the teams. Triggered by game events like kills or just simple mistakes, players begin to turn sour. The communication channels that were meant to coordinate the team effort can then be used to verbally assault other players, often by using profane terms and heavy insults.

Collecting bad game experiences like this is harmful for the community, as it can bias a player's attitude towards engaging in cooperation even when confronted with fresh opponents and new teammates in later matches. The perceived hostility in a player community is frequently referred to as *toxicity*. Toxicity imposes a serious challenge for game designers, as it may chase active regular players away. It might also prevent new players from joining the game, because a toxic base appears as unfriendly and hostile to newcomers.

The main contribution of this work is to devise an annotation system for chats of multiplayer online games that can be used for detecting toxicity (Section III). We apply the system to a large dataset (Section II) collected from a representative game of the MOBA genre and propose a method based on machine learning that uses the annotation system to predict the outcome of ongoing matches (Section IV). We end with related work (Section V) and conclusions (Section VI).

II. Data

A. Data sources

All data used in this work are based on one of the ancestors of all MOBA games: Defense of the Ancients (DotA).[4] This game started as a custom map for the real-time strategy game Warcraft III but soon became so popular that community platforms emerged that allowed for players to register, get profiled and being matched up against each other based on their skill. One of these platforms was DotAlicious, from which we crawled our data.

The website of DotAlicious is no longer available online, as DotA has been substituted by newer MOBAs like League of Legends, Heroes of the Storm or Dota II. The core game principles have not been changed by much by DotA's successors, but the accessibility of replays, chat-logs and player-related information for them is more limited due to several privacy

M. Martens, S. Shen, A. Iosup, and F. Kuipers, "Toxicity detection in multiplayer online games," Annu. Work. Netw. Syst. Support Games, vol. 2016-Janua, 2016

Model prediksi dibangun menggunakan **Support Vector Machine (SVM)** dan feature menggunakan **TF-IDF** dengan **N-gram**.

Penelitian tersebut bertujuan untuk menunjukkan bahwa perilaku toxic memiliki dampak, sehingga riwayat percakapan dapat dijadikan fitur untuk memprediksi hasil dari pertandingan dalam game tersebut.

Penelitian tersebut menemukan bahwa tim yang kalah lebih cenderung perilaku toxic. Pada penelitian ini peneliti mencoba mengambil bagian bad word saja untuk Memprediksi hasil pertandingan dan **akurasinya rendah yaitu sekitar 65%**.

Kesimpulannya penggunaan **bad word saja tidak bisa diandalkan** untuk memprediksi Hasil pertandingan.

	t = 0.5			t = 0.75			t = 1.0		
	#features	avg accuracy	std accuracy	#features	avg accuracy	std accuracy	#features	avg accuracy	std accuracy
all words	127612	0.6399	0.0140	170063	0.7689	0.0092	208598	0.9407	0.0048
all but "command"	126900	0.6346	0.0103	169298	0.7421	0.0099	207758	0.8708	0.0070
only "bad"	1442	0.5720	0.0137	1767	0.6077	0.0096	2020	0.6538	0.0108
only "slang"	880	0.5877	0.0189	908	0.6875	0.0114	921	0.8295	0.0093

- Model identifikasi toxic menggunakan rule based
- n-gram yang sudah dibangun dipilih sebanyak 100 n-gram token yang kemunculannya terbanyak
- Untuk 1-gram proses penentuan berdasarkan token tersebut memiliki makna menghina atau tidak
- untuk 2-gram, 3-gram, dan 4-gram token tersebut memiliki makna menghina ke seseorang atau tidak
- N-gram token yang telah dipilih sebagai toxic dibandingkan dengan kalimat yang ingin diidentifikasi

Named Entity Recognition

Pembangunan NER dengan Conditional Random Fields (CRFs) yang umum digunakan untuk labeling dan parsing sequential data.

TEKNIK

Feature Extraction

- Word parts
- Lower/title/upper flags
- Features of nearby words)
- Ubah menjadi format sklearn-crfsuite
- Setiap kalimat dikonversi ke daftar kamus.

Lalu dilakukan split data untuk train (70%) dan test data (30%)

DATA

Train model menggunakan dataset dota 2 chat sebanyak 241 kalimat yang sudah dilakukan manual labeling untuk Entity Tags (O, hero, praise, bad). Dengan format tagging IOB. Bentuk datasetnya sebagai berikut:

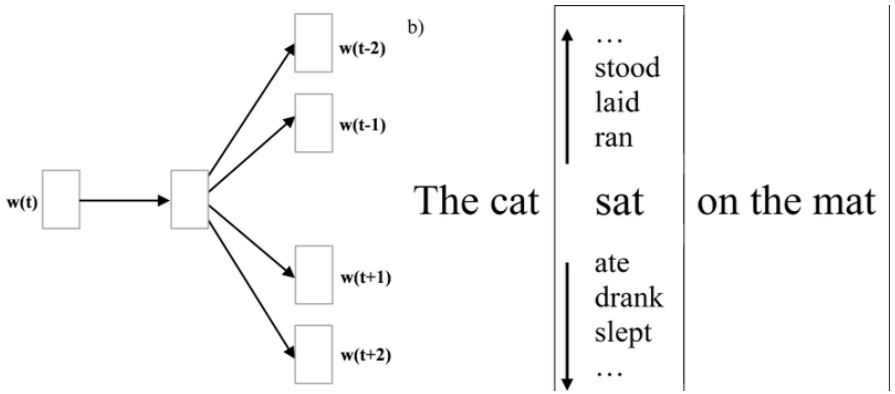
Kalimat	Kata	Tag
Sentence: 1	yes	O
	dog	B-bad
Sentence: 2	yeah	O
Sentence: 3	fast	O
	and	O
	furious	O
Sentence: 4	too	O
	fas	O
Sentence: 5	haha	O
Sentence: 6	sad	O
...	...	...

HASIL

Precision	Recall	F1-Score	Support
B-bad	1.00	0.25	0.40
B-her	0.00	0.00	0.00
B-pra	0.67	0.33	0.44
I-her	0.00	0.00	0.00
I-bad	1.00	0.50	0.67
I-pra	0.75	0.38	0.50
O	0.86	0.98	0.92
accuracy			0.85
macro avg	0.55	0.32	0.38
weighted avg	0.81	0.85	0.81

Word Embedding

Pembangunan model Word Embedding menggunakan teknik Skip-Gram.



Source : https://miro.medium.com/max/1700/0*yxs3JKs5bKc4c_i8.png

TEKNIK

Membuat word index table:

Indeks	Token
1	OOV
2	i
3	gg
...	...

Membuat Tabel word embedding :

Indeks	Vector Word Embedding
1	[0.00000000e+00, 0.00000000e+00, ..., 0.00000000e+00]
2	[-0.25935695, -0.24364145, ..., 0.35460544]
3	[-2.42174849e-01, -2.28570521e-01, ..., 3.44555348e-01]
...	...

DATA

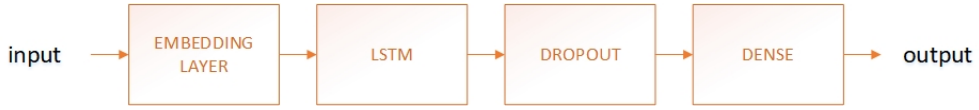
Bentuk datasetnya sebagai berikut:

Dokumen	Kalimat Ke	Text
1	1	yes dog
1	2	yeah
1	3	fast and furious
1	4	too fas
1	5	haha
1	6	sad
2	1	no idiot
2	2	we too pro
2	3	lol
...	...	...

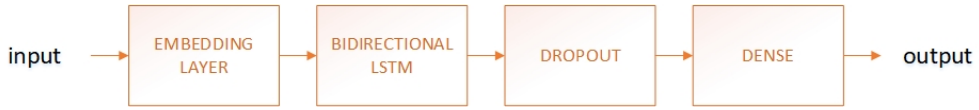
Text Classification

TEKNIK

1. LSTM



2. Bidirectional LSTM



Input: Daftar token dari dokumen yang telah di encoding menggunakan tabel word index.

Output: Hasil classification berupa nilai peluang dokumen tersebut terhadap suatu kelas.

DATA

Bentuk datasetnya sebagai berikut:

Category	Match	Slot	Text
0	2	0	yes dog. yeah . fast and furious. too fas. haha. sad.
0	2	2	no idiot. we too pro. lol.
...	...	...	

Penjelasan Kolom:

- Category: Jika value 0 berarti tidak toxic dan 1 berarti toxic.
- Match: ID pertandingan
- Slot: slot pemain pada pertandingan tersebut
- Text: Kumpulan percakapan pemain pada pertandingan tersebut.

Text Classification

DATA

Jumlah Category dalam dataset sebagai berikut:

EKSPERIMEN

Eksperimen akan dilakukan dalam beberapa skenario untuk mencari model terbaik dengan menggunakan validation test dan dengan menggunakan beberapa parameter yaitu:

1. Teknik: LSTM dan Bidirectional LSTM
2. Entity Masking: Memakai Entity Masking dan tanda Entity Masking
3. Panjang Vector Embedding: 50 dan 100
4. Dropout Layer: Memakai Dropout Layer dan tanpa Dropout Layer

Text Classification

SKENARIO

Skenario eksperimen untuk pencarian model terbaik sebagai berikut:

Skenario	Teknik	Entity Masking	Panjang Vector Embedding	Dropout Layer
A	LSTM	Tidak	50	Tidak
B	LSTM	Tidak	50	Ya
C	LSTM	Tidak	100	Tidak
D	LSTM	Tidak	100	Ya
E	LSTM	Ya	50	Tidak
F	LSTM	Ya	50	Ya
G	LSTM	Ya	100	Tidak
H	LSTM	Ya	100	Ya
I	Bidirectional LSTM	Tidak	50	Tidak
J	Bidirectional LSTM	Tidak	50	Ya
K	Bidirectional LSTM	Tidak	100	Tidak
L	Bidirectional LSTM	Tidak	100	Ya
M	Bidirectional LSTM	Ya	50	Tidak
N	Bidirectional LSTM	Ya	50	Ya
O	Bidirectional LSTM	Ya	100	Tidak
P	Bidirectional LSTM	Ya	100	Ya

Data train dan data test akan dibagi menjadi **70 : 30**. Ratio ini dipilih karena order datasetnya hanya ratusan.

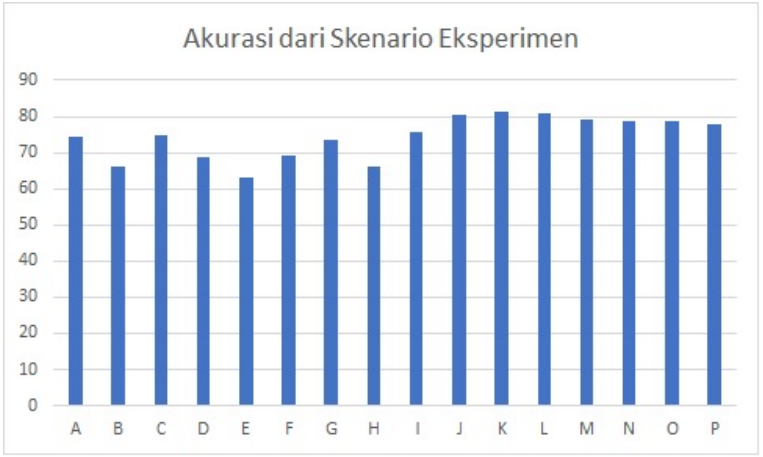
Menggunakan **5-Fold Validation** dan Metric yang digunakan sebagai alat ukur performa model menggunakan **F1 Score**, karena jumlah category 0 dan 1 tidak seimbang.

HASIL EKSPERIMEN

VALIDATION

Berikut hasil validation test:

Skenario	F1 Score (%)
A	74.21
B	66.29
C	74.83
D	68.69
E	63.24
F	69.11
G	73.44
H	66.01
I	75.81
J	80.57
K	81.19
L	80.91
M	79.30
N	78.87
O	78.76
P	77.72



Sehingga Model Terbaik yang dipilih pada **Skenario K**

TESTING

Model terbaik digunakan untuk diuji dengan data test dan hasil **71.42%**

ANALISIS

1 Model terbaik, cenderung menganggap dokumen tersebut toxic jika dokumen mengandung kata yang banyak.

DOKUMEN 1

what. jeje fAM. free farming Is. not coming into play. let end. storm fat yet. zZZ. ok. U useless anyways. does it matter. 30mins in. cant seem to hit a singel call. ROFL. still didnt hirt. Aha. better share hero contorl. to someone else. might start hitting ur Q. ROFL. fuckING. Retard. yea. he had an. amazing blast laen. at mid. 20mins scythe on od. Yet my . LS . wants to farm. His orchid. Die btich. oh. my god. oh my god. oh. My. Fucking god (Berlabel toxic)

DOKUMEN 2

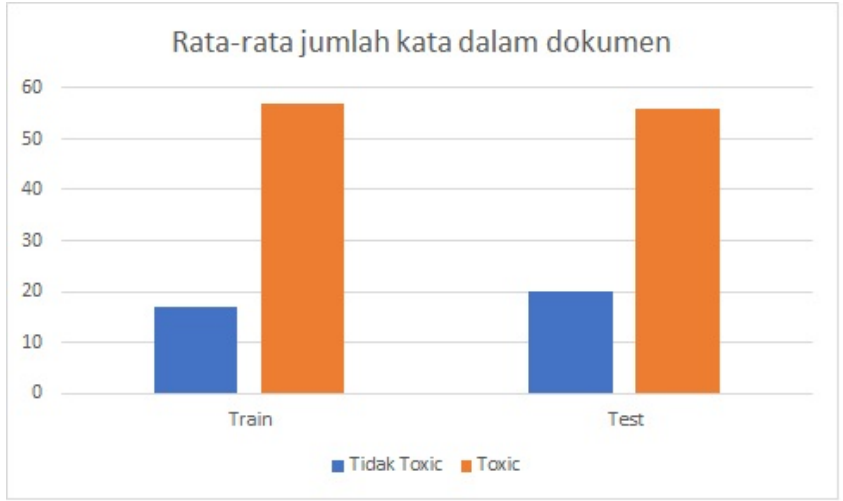
what's happening boyz ?. as you can see we are waiting . :D. I am just asking what happened. <3. how many more x3 mins do we have to wait. :D. BOI. you don't know how time works. look. you will have more gold if he leaves. If I roll a one. WE GO. izi. +. now you are lucky. :D. naah we ain't. :D. gg wp (Berlabel tidak toxic)

DOKUMEN 3

so ya mama likes dick ehh?. figures. ur not even a good hooker kid. passive shadow blade?. gg (berlabel toxic)

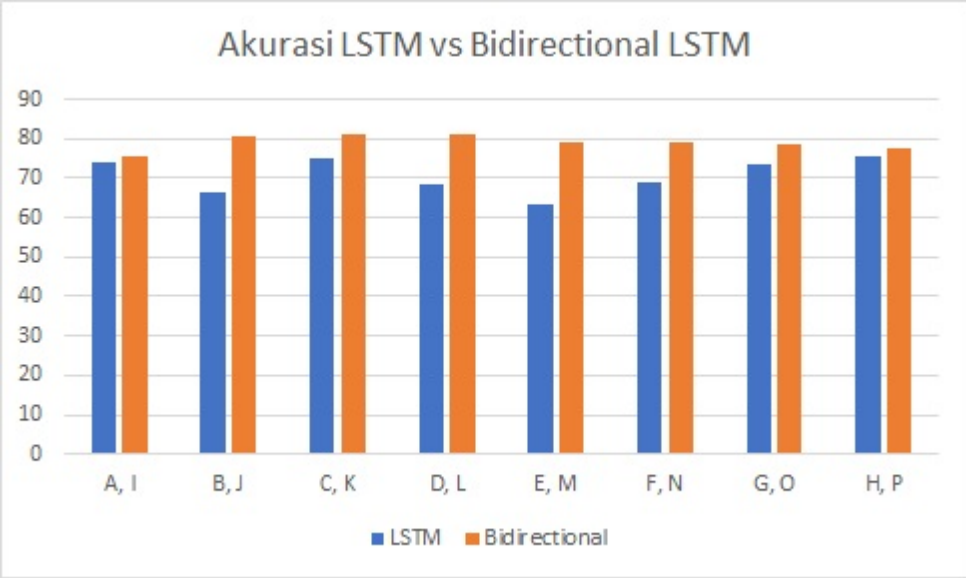
DOKUMEN 4

drow . remember me last game. haha. waot. haha (berlabel tidak toxic)



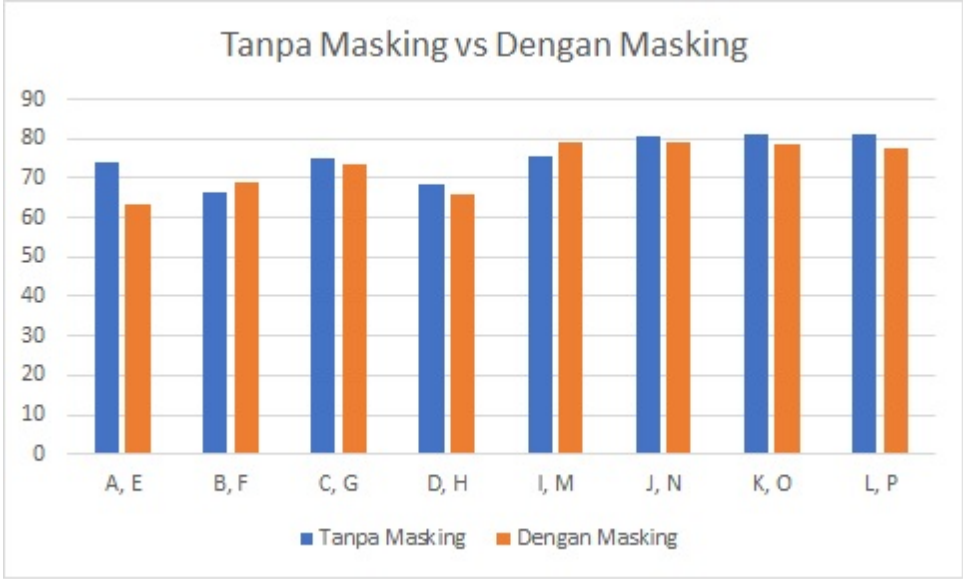
ANALISIS

2 Model bidirectional memiliki akurasi yang lebih baik dibanding dengan model LSTM. bidirectional mempertimbangkan hubungan kata dengan kata sebelum dan setelahnya.



ANALISIS

3 Penggunaan Entity Masking tidak terlalu mempengaruhi akurasi. disebabkan banyak entity selain OTHER yang gagal diidentifikasi. tergambar di akurasi untuk label selain other sangat rendah dengan support rendah. Solusi, menambah dataset untuk meningkatkan akurasinya.



Contoh:

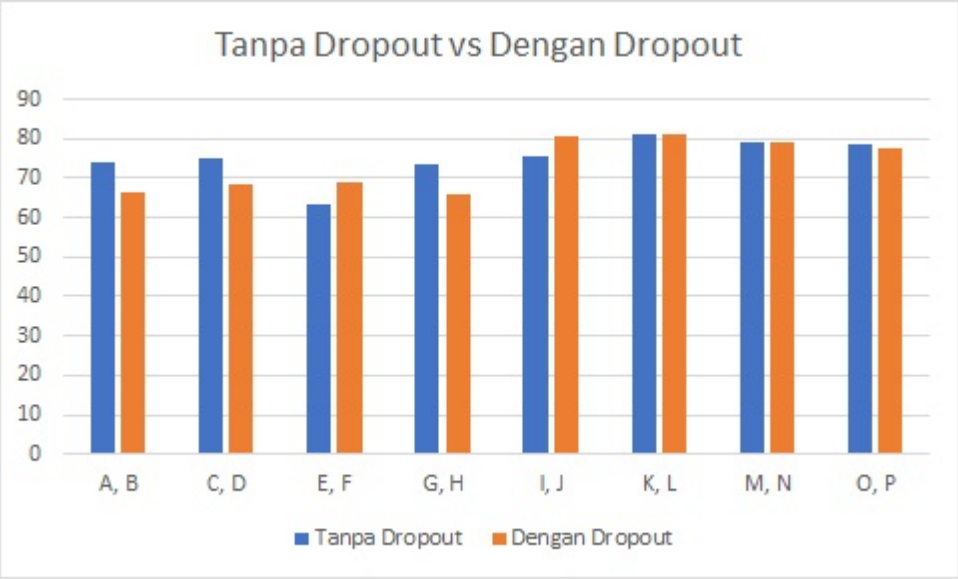
DOKUMEN

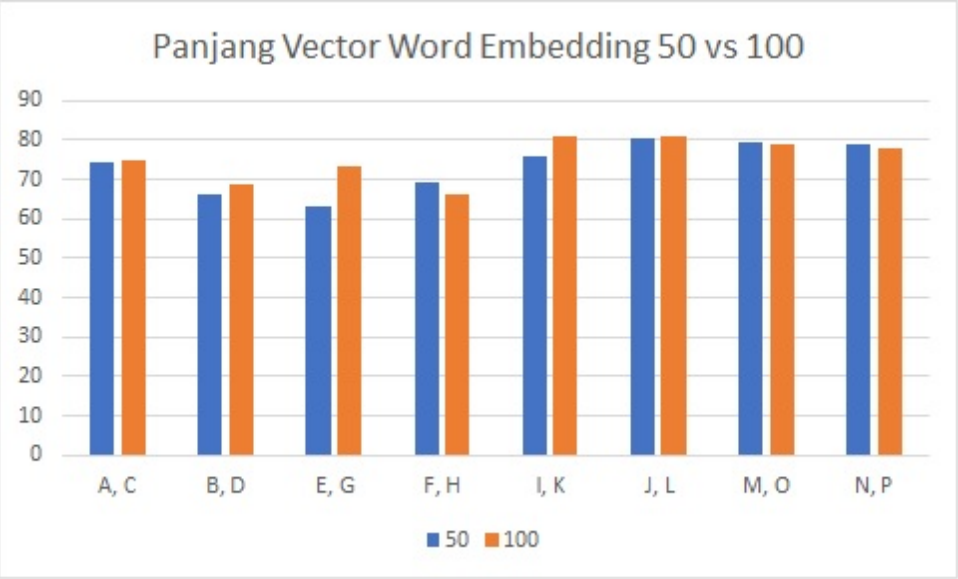
'die idoot . divine for the win . so noob . fucker . Tk . ahahaha . so sad . triggerd . rep [ort jugg . ty . ez . just erport this jugg . ty . gg'.

Tidak ada masking yang terindentifikasi pada document tersebut, padahal kata idoot, noob, fucker bermakna BAD.

ANALISIS

4 Dua parameter lainnya, yaitu panjang vector word embedding dan penggunaan dropout layer atau tidak, tidak terlalu mempengaruhi hasil. Sebagai berikut:





REFERENSI

1. “Why is Dota 2 dying? - Quora.” [Online].
Available: https://www.quora.com/Why-is-Dota-2-dying. [Accessed: 14-Mar-2020].

2. “Is DotA Dying? :: Dota 2 General Discussions.” [Online].
Available: https://steamcommunity.com/app/570/discussions/0/1744483505474625833/. [Accessed: 14-Mar-2020].

3. “4 Alasan Mengapa Kini Dota 2 Mati. - Gamebrott.com.” [Online].
Available: https://gamebrott.com/4-alasan-mengapa-kini-dota-2-mati. [Accessed: 14-Mar-2020].

4. M. Martens, S. Shen, A. Iosup, and F. Kuipers,
“Toxicity detection in multiplayer online games,” Annu. Work. Netw. Syst. Support Games, vol. 2016-Janua, 2016.

Terimakasih

Question and Answer

Email:
23519009@std.stei.itb.ac.id
23519035@std.stei.itb.ac.id