Implementasi Deep Learning untuk mengubah kalimat tidak sopan menjadi sopan

Axel Christiant
Program Studi Informatika – Program Sarjana
Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman,
Yogyakarta, Indonesia
axel.christiant@students.uii.ac.id

Ahmad R. Pratama
Jurusan Informatika
Universitas Islam Indonesia
Jl. Kaliurang KM 14,5 Sleman,
Yogyakarta, Indonesia
ahmad.rafie@uii.ac.id

***Abstract*—Menurut survey yang dilakukan oleh Microsoft pada tahun 2020, Indonesia menjadi negara dengan kesopanan digital paling buruk di Asia Pasifik. Hal tersebut dibuktikan dengan naiknya angka Digital Civility Index 8 poin dari tahun 2019 menjadi 76 poin. Oleh sebab itu, diperlukannya sistem yang dapat membantu masyarakat Indonesia untuk dapat berkomunikasi secara sopan di dunia digital. Hal tersebut menginspirasi penelitian ini untuk membangun sistem deep learning untuk mengubah kalimat menjadi sopan. Metode yang digunakan penelitian ini adalah Tag and Generate Approach; model tagger untuk menggantikan token tag pada kata yang terdapat di kalimat tidak sopan dan model generator untuk menggantikan token tag tersebut dengan kata yang sesuai sehingga menjadi kalimat yang sopan. Sebelum melakukan pelatihan model tagger, setiap n-gram (penelitian ini menggunakan jangkauan unigram sampai bigram) dilakukan penghitungan rasio tf-idf untuk mengetahui peringkat persentil dan relevansi n-gram pada masing-masing gaya teks. Dapat disimpulkan bahwa kata "mengapa" dan "kalo" menduduki peringkat persentil unigram tertinggi, sedangkan kata "bagaimana ini" dan "ya min" mendapatkan peringkat persentil tertinggi bigram pada masing-masing gaya teks. Hasil akhir evaluasi model mencapai nilai tertinggi pada BLEU 1 dengan nilai 53.48 dan disusul oleh BLEU 2 sebesar 40.30. Sedangkan, untuk metrik BLEU 3, BLEU 4, dan METEOR masih tertinggal jauh dibandingkan kedua metrik tersebut (BLEU 1 dan BLEU 2) dengan nilai masing-masing 31.2, 24.5, dan 26.2. Berdasarkan hasil tersebut, dapat disimpulkan bahwa model mencapai skor tertinggi pada content preservation pada unigram dan bigram.**

***Keywords*—Text Style Transfer, Politeness, Tag and Generate, Deep Learning**

I. PENDAHULUAN

Kesopanan adalah perilaku yang mempertimbangkan perasaan orang lain tentang bagaimana mereka harus diperlakukan secara interaksional yang menunjukkan kepedulian terhadap status dan hubungan sosial dari para pelaku interaksi [1]. Definisi tersebut menunjukkan bahwa kesopanan merupakan salah satu hal yang krusial dalam membangun dan mempertahankan hubungan sosial antar manusia.

Kemajuan teknologi membuat manusia semakin mudah dalam melakukan komunikasi. Namun di sisi lain, hal ini juga menyebabkan manusia menjadi merasa bebas dari norma kesopanan karena tidak adanya pengawasan langsung dari pihak otoritas. Hal tersebut dapat dibuktikan dengan hasil survey dari Microsoft yang memiliki parameter bernama *Digital Civility Index* untuk mengukur tingkat adab atau kesopanan masyarakat di setiap negara saat berselancar di dunia digital. *DCI* mempunyai skala dari 0 sampai 100 poin yang semakin tinggi nilainya maka semakin buruk tingkat kesopanan masyarakat di negara tersebut. Indonesia menjadi negara dengan kesopanan digital paling buruk di Asia Pasifik, hal ini dapat dibuktikan dengan angka *DCI* bernilai 76 pada tahun 2020 yang naik 8 poin dari tahun 2019 [2] . Nilai tersebut menunjukkan bahwa masyarakat Indonesia belum dapat melaksanakan norma kesopanan dengan baik di dunia digital terutama masyarakat dengan usia dewasa dan remaja dengan kontribusi 83% dan 68%.

Ketidaksantunan dalam berbahasa terjadi karena adanya penggunaan tuturan yang informal dalam situasi yang formal (adanya jarak sosial) atau sebaliknya [3]. Oleh karena itu, konteks perlu digunakan dalam memahami dan menghasilkan tuturan untuk membangun kerjasama dan sopan santun dalam proses komunikasi sehingga tujuan komunikasi dapat dicapai secara efektif [4]. Meskipun banyak orang menyadari pentingnya kesantunan berbahasa dalam komunikasi bukan berarti hal ini mudah dilakukan di dunia digital. Hal ini dapat disebabkan karena kurang kritisnya seseorang akan tuturan dalam berkomunikasi, kurangnya pengetahuan dalam kesantunan berbahasa, dan butuhnya menyampaikan tuturan secara cepat dalam berkomunikasi sehingga tidak sempat mengubahnya menjadi bahasa yang sopan. Oleh karena itu, dalam berkomunikasi di media sosial dibutuhkanlah sistem yang dapat membantu seseorang untuk berkomunikasi secara sopan di platform digital.

Penelitian ini akan membangun sistem *NLG (Natural Language* Generation*)* berjenis *Text Style Transfer* untuk mengubah kalimat tidak sopan menjadi kalimat yang sopan menggunakan strategi mempertahankan bentuk formalitas pada kalimat dengan cara mengubah kata yang tidak baku menjadi baku [5], sebagai contoh :

Kalimat tidak sopan: min, aq pengen main mobile legend kok tp lag banget yaaa.

Kalimat sopan: Admin, saya ingin bermain mobile legend tetapi mengapa lag sekali.

Metode *TST* yang digunakan pada penelitian ini adalah *Tag and Generate Approach.* Metode tersebut digunakan pada penelitian sebelumnya yang berjudul *Politeness Transfer : A Tag and Generate Approach* [6] untuk mengubah data kalimat percakapan email karyawan *Enron* menjadi kalimat yang sopan.

Gambar 1. Implementasi pipeline Tagger dan Generator [5]

II. KAJIAN PUSTAKA

Gaya teks (*Text Style*) adalah variasi linguistik pada teks dengan tetap mempertahankan konteks yang ingin disampaikan. Dalam melakukan komunikasi, gaya teks yang digunakan setiap orang merupakan sesuatu yang situasional dan tidak pasti. Setiap kata penyusunnya didasarkan pada waktu, tempat, dan skenario tertentu untuk menyampaikan maksud dari pembicara dengan karakteristiknya sendiri [7].

Secara intuitif, gaya pada teks mendeskripsikan tentang bagaimana cara komunikator memilih kata yang dipakai dan menyusun kalimat yang dipakai untuk membentuk intonasi, gambar, dan semantik pada teks [8].

Beberapa tahun terakhir, penelitian tentang gaya pada teks atau *Text Style* menarik perhatian tidak hanya ahli bahasa namun juga ilmuwan komputer. Secara spesifik, penelitian yang dilakukan oleh ilmuwan komputer adalah transfer gaya pada kalimat atau disebut juga dengan *Text Style Transfer (TST). Text Style Transfer* adalah bagian cabang dari *Natural Language Generation (NLG)* yang bertujuan untuk mengubah suatu gaya pada suatu teks dengan tetap mempertahankan konten pada teks tersebut.

Pada awal perkembangannya, *TST* masih berkaitan erat dengan *Neural Machine Translation (NMT)* dan *Neural Style Transfer (NST).* Hal tersebut dikarenakan *NMT* dan *NST* juga merupakan suatu cabang dari *Natural Language Generation* yang bertujuan untuk melatih komputer menghasilkan teks yang dapat dipahami oleh manusia berdasarkan input dari data yang ada. Untuk menggambarkan suatu implementasi dari *Text Style Transfer*, diasumsikan terdapat dua jenis teks yaitu: formal (*x*) dan informal (*x'*). Tugas sebuah model *TST* adalah menerima *input x* dengan atribut dasar teks *s* dan model tersebut akan menghasilkan *output x'* menggunakan attribut teks pada target *t* dengan tetap mempertahankan konten dan konteks pada *x* [8].

Terdapat dua jenis data yang digunakan untuk melatih model *Text Style Transfer* yaitu paralel dan non-paralel. Data berjenis paralel berisi sepasang teks yang masing-masing memiliki konteks yang sama namun gaya penulisan teks yang berbeda. Berlawanan dengan non-paralel, data berjenis non-paralel mempunyai sepasang teks yang berbeda konteks dan gaya penulisan teks. Dataset berjenis paralel membutuhkan lebih banyak pekerjaan manusia dalam mengubah gaya pada satu teks ke gaya lain karena data yang terdapat di dunia nyata hampir tidak ada yang tersedia secara paralel.

Untuk melakukan evaluasi pada model *Text Style Transfer* terdapat tiga objektif untuk mengukur kualitas suatu model:

1. ***Content Preservation*** atau kelestarian konten pada teks yang dihasilkan oleh model tidak berubah terhadap input. **BLEU** [9] dan **METEOR** [10] adalah metrik yang dibuat untuk mengukur kualitas dari *Machine Translation* atau sistem penerjemah bahasa. Kedua metrik tersebut menghitung skor berdasarkan kesamaan antara teks yang dihasilkan oleh model dan referensi teks yang dibuat untuk evaluasi model. **BLEU** mengukur teks pada *output* dengan menghitung kesamaan pada *n—grams* atau jumlah kata yang sama pada kedua teks (*output* dan teks *reference*) menggunakan *precision* [6] sedangkan **METEOR** adalah metrik hasil dari modifikasi *weighted F-Score* berdasarkan pencocokan *unigram* dan *penalty function* untuk kata yang tidak sesuai urutan.
2. ***Quality of style*** atau kualitas model untuk menghasilkan gaya teks target pada *output.* Untuk mengukur kualitas gaya teks yang dihasilkan oleh model diperlukannya *classifier* untuk menghitung persentase akurasi teks tersebut terhadap gaya teks yang dituju. Kebanyakan pada penelitian *Text Style Transfer* sebelumnya [11][12][13] menggunakan *pre-trained classifier* untuk mengukur akurasi keakuratan gaya teks model mereka. Semakin tinggi akurasi, kualitas gaya teks yang dihasilkan model. Metrik berupa *precision, recall, dan F1* score juga dapat mengukur kualitas model dalam hal tersebut.
3. ***Fluency*** atau kefasihan bahasa pada teks yang dihasilkan oleh model untuk dapat dipahami oleh manusia secara kontekstual. *The Kneser-Ney language model* [14] adalah *pre-trained model* yang biasanya digunakan untuk mengukur kualitas kefasihan bahasa pada model. *The Kneser-Ney language model* mengukur *perplexity score* dari teks yang dihasilkan oleh model dengan cara membandingkan *trigram* pada teks dengan distribusi *trigram* yang telah diestimasi pada pelatihan model. Semakin kecil *perplexity score* yang didapat maka semakin fasih model *TST* tersebut [8].

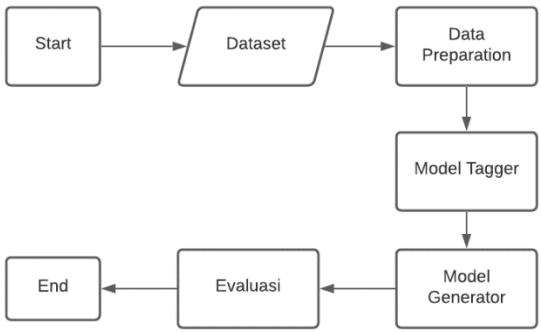
Metode *Tag and Generate Approach* pada *Text Style Transfer* dibuat oleh Aman Madaan [6] dengan tujuan untuk mengubah kalimat yang tidak sopan menjadi kalimat yang sopan menggunakan data email karyawan *Enron.* Metode tersebut dibagi menjadi dua tahap, yaitu:

1. ***Tagger*****:** Mengidentifikasi kata atau frasa yang merupakan suatu atribut pada gaya teks tersebut lalu mengubahnya menjadi token tag.
2. ***Generator*****:** Menerima input dari *tagger* dan mengisi *token tag* dengan atribut pada gaya teks target. Jika suatu input sudah tidak memiliki *token tag* maka teks tersebut sudah memenuhi atribut pada gaya teks target dan tahap ini menghasilkan teks yang sama dengan *input*.

Gambar 1 menunjukkan implementasi *pipeline Tag and Generate approach.* Pada contoh pertama (***add - tagger***), teks $\mathbf{x_1^{(1)}}$ tidak mempunyai atribut gaya teks yang menjadi karateristik teks tersebut, dengan begitu *tagger* menambahkan *token tag* tanpa harus menghapus kata atau frasa yang terdapat pada teks z(x1). Pada contoh kedua (***replace - tagger***), teks $x_2^{(1)}$ mempunyai atribut teks berupa kata *ok* dan *bland* yang menjadi karateristik gaya teks *sentiment negative* karenanya *tagger* menghapus dan mengisi kata tersebut dengan *token tag* pada teks z(x2). Setelah teks diproses oleh tahap *tagger*, output pada model tersebut menjadi input untuk model *generator* menggantikan *token tag* menjadi kata atau frasa yang memiliki karakteristik gaya teks pada target.

Hasil yang didapatkan pada metode ini jika dibandingkan dengan metode penelitian sebelumnya [11] dengan dataset yang sama; mengalami peningkatan sebesar 58.61 pada metrik BLEU, 18.07 pada metrik METEOR, dan 0.7 pada *Content preservation.*

III. METODOLOGI PENELITIAN



Gambar 2. Diagram alur penelitian

A. Dataset

Penelitian ini menggunakan dataset yang telah digunakan pada penelitian sebelumnya dengan judul "*Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation"* [15]. Dataset berisikan total 5000 teks (2500 teks tidak sopan dan 2500 teks sopan). Dataset tersebut memenuhi kriteria batasan masalah pada penelitian ini yaitu menggunakan strategi kesopanan dalam komunikasi dengan cara mempertahankan bentuk formalitas pada kalimat yang disampaikan [5]. Kumpulan teks yang terdapat pada dataset diambil 2500 dari 52.000 *tweet* yang masih berbentuk teks informal lalu dibuatlah bentuk formal pada masing-masing teks tersebut. Karakteristik teks tidak sopan pada dataset dapat dijelaskan sebagai berikut:

1. **Menggunakan kata sehari-hari atau mempersingkat kata.**

   Contoh: gw knp tdk bs.

   Diubah menjadi: Saya kenapa tidak bisa?

2. **Pengunaan afiks dan sufiks yang tidak sesuai.**

   Contoh: saya mesenin taxi.

   Diubah menjadi: saya memesan taxi.

3. **Urutan kata yang kurang tepat**

   Contoh: Admin, bisa seperti itu kenapa?

   Diubah menjadi: Admin, kenapa bisa seperti itu?

B. Data-preparation

Sebelum melakukan pelatihan model *tagger dan generator* dibutuhkan persiapan dan pemrosesan data agar dapat dilatih dengan baik di model *tagger*. Persiapan data pertama kali yang dilakukan adalah mengubah label menjadi P_9 untuk teks berjenis sopan dan P_0 untuk teks netral atau tidak sopan. Selanjutnya, memberikan *tag token* berdasarkan estimasi setiap frasa pada masing masing gaya teks menggunakan persamaan (1) dan (2).

$$n_1^2(w) = \frac{\frac{1}{m}\sum_{i=1}^{m} tf - idf\left(w, x_i^{(2)}\right)}{\frac{1}{n}\sum_{j=1}^{n} tf - idf\left(w, x_j^{(1)}\right)} \quad (1)$$

$$p_1^2(w) = \frac{n_1^2(w)^\gamma}{\sum_{w'} n_1^2(w')^\gamma} \quad (2)$$

Diasumsikan diberikan pasangan korpus X_1 dan X_2 dengan gaya teks masing-masing S_1 dan S_2 dan w merupakan *sampling n-grams* yang diambil dari kedua korpus, maka $p_1^2(w)$ menghitung distribusi probabilitas berdasarkan kemunculan kata pada kedua korpus. Secara intuitif, $p_1^2(w)$ merupakan persamaan yang proposional terhadap probabilitas kemunculan suatu *n-gram* pada kedua korpus namun memiliki nilai *tf-idf* yang lebih tinggi pada X_2 dibandingkan X_1. Sedangkan, $\eta_1^2(w)$ merupakan rasio dari rerata *tf-idf* pada suatu *n-gram* terhadap korpus X_1 dan X_2. *N-gram* dengan nilai $\eta_1^2(w)$ yang tinggi memiliki rerata tf-idf yang tinggi pada korpus X_2 dibanding X_1, yang berarti kata tersebut memiliki karakteristik gaya teks S_2 [6].

C. Training Tagger dan Generator

Dalam melakukan *training,* model *tagger* menerima *input* teks yang berasal dari data *training* sedangkan model *generator* menerima *input* dari *output* pada model *tagger.* Model *tagger* memiliki dua varian model yang setiap variannnya bergantung pada tujuan yang ingin dicapai. Varian pertama ***replace-tagger*** merupakan model yang mengidentifikasi attribut gaya teks pada *input* $\alpha\left(x_1^{(1)}\right)$ lalu menggantikannya dengan *TAG token;* varian ini tepat untuk model yang mempelajari atribut dari kedua gaya teks (*source* dan *target*), seperti misalnya *sentiment transfer* yaitu membutuhkan model untuk mempelajari atribut pada sentimen positif dan negatif. Sedangkan, varian kedua ***add-tagger*** merupakan model yang mengidentifikasi posisi yang tepat untuk menambahkan *token tag* tanpa menghapus kata atau frasa yang ada; varian ini yang dipakai pada penelitian sebelumnya [6]. Sebab untuk mengubah kalimat netral menjadi kalimat yang sopan, model hanya perlu mendapatkan informasi tentang atribut dari gaya teks sopan. Secara umum, kedua varian model tersebut memiliki tujuan yang sama yaitu memberikan *tag token* pada teks lalu nantinya akan menjadi *input* untuk model *generator.*

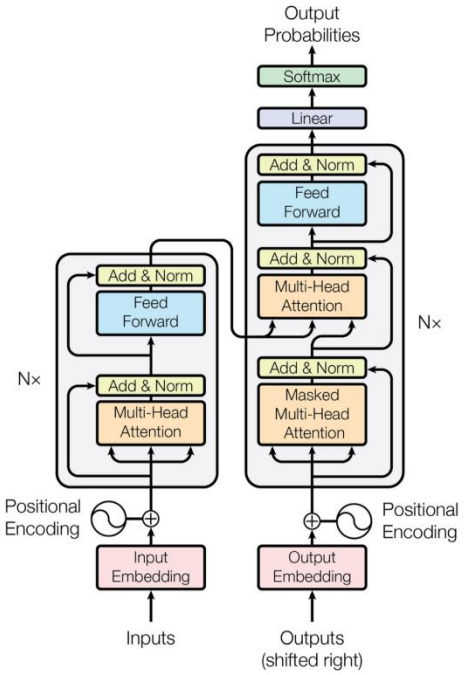
$$L_r(\theta_t) = -\Sigma_{i=1}^{|x_1|} log p_{\theta_t}\left(z(x_i)|x_i^{(1)}; \theta_t\right) (3)$$

$$L_a(\theta_t) = -\sum_{i=1}^{|x^1|} log p_\theta t \left(z(x_i)|x_i^{(2)} \backslash a(x_i^{(2)})\right) (4)$$

Persamaan (3) adalah *loss function* yang digunakan untuk melatih model varian *add-replace* dan persamaan (4) untuk model varian *add-tagger*. Pada penelitian ini, model *tagger* membutuhkan informasi atribut pada teks tidak sopan. Oleh sebab itu, model varian *add-replace* yang akan dipakai untuk melatih model *tagger* dan menggunakan persamaan (3) sebagai *loss function.*

$$L(\theta_g) = -\Sigma_{l=1}^{|x_v|} log p_{\theta g}\left(x_i^{(v)}|Z(x_i); \theta g\right) (5)$$

Sedangkan persamaan (5) merupakan *loss function* yang akan dipakai untuk melatih model *generator*. Pada pelatihan model *generator* digunakan juga teknik *data augmentation* berupa *random word shuffle* dan *word drops replacement.* Model *tagger* dan *generator* menggunakan 4 lapis model arsitektur *transformer*, di setiap model *transformer* tersebut memiliki 4 *attention heads* dengan 512 *dimensional embedding* dan *hidden state size.* Ditambahkan juga *Dropout* dengan nilai p = 0.3 pada setiap lapisan *transformer*.



Gambar 3. Arsitektur *transformer*

D. Evaluasi

Untuk melakukan evaluasi terhadap performa model, penelitian ini terbatas pada *content preservation* dan *fluency* dari teks yang dihasilkan. Metrik evaluasi yang digunakan untuk mengukur *content preservation* pada penelitian ini adalah **BLEU [9]** dan **METEOR [10].**

BLEU merupakan metrik yang menghitung jumlah kesamaan *n-gram* pada teks yang dihasilkan model dengan teks pada target. Untuk mendapatkan hasil akhir **BLEU**, hal pertama yang dilakukan adalah menghitung rata-rata geometrik *modified precision score* pada korpus yang dilakukan pengujian (p_n) menggunakan *n-grams* dengan batas N dan bobot (w_n) yang jika dikumulatifkan bernilai 1. Selanjutnya dikalikan oleh *exponential brevity penalty factor* (BP). Seperti yang dapat dilihat pada persamaan (6).

$$BLEU = BP.\exp(\Sigma_{n=1}^{N} w_n log p_n) (6)$$

Secara singkat, **BLEU** mempunyai skala dari 0 sampai 1. Semakin mendekati 1 nilainya maka semakin banyak kesamaan *n-gram* antara teks yang dihasilkan model dengan teks target. Evaluasi yang dilakukan pada penelitian ini menggunakan **BLEU** dengan jangkauan *n-gram*: *unigram* sampai dengan 4-*gram* atau **BLEU 1** sampai dengan **BLEU 4**.

Pada metrik kedua yaitu **METEOR,** metrik tersebut dianggap lebih baik daripada **BLEU** saat dibandingkan menggunakan parameter penilaian manusia. **METEOR** merupakan modifikasi dari perhitungan *precision* dan *recall* dengan *weighted F-Score* berdasarkan pemetaan *unigram* pada teks yang dihasilkan model dengan teks target.

$$F = \frac{PR}{\alpha P + (1 - a)R} (7)$$

Untuk menghitung *weighted F-Score*, yang pertama kali dilakukan adalah menghitung jumlah pemetaan (m) *unigram* pada kedua teks (teks yang dihasilkan oleh model dan teks target). Setelah itu dilakukan penghitungan *precision: m/c* dan *recall: m/r*, yang mana c adalah panjang teks yang dihasilkan model dan r adalah panjang teks pada target. Hasil akhir *weighted F-Score* dapat dihitung menggunakan persamaan (7).

Berbeda dengan **BLEU, METEOR** menghitung *penalty* berdasarkan urutan *n-gram* pada teks yang dihasilkan model tidak sesuai dengan teks target. Hasil akhir dari skor **METEOR** didapatkan melalui (1 - *Penalty*)*F* dengan jangkauan *range* skor dari 0 sampai 1. Dapat disimpulkan bahwa semakin tinggi skor **METEOR** maka semakin banyak kesamaan *n-gram* juga teks yang dihasilkan oleh model dengan teks pada target (berdasarkan kecocokan *n-gram* dan urutannya).

IV. HASIL DAN PEMBAHASAN

.

A. Hasil Data Preparation

Data preparation memiliki peran penting untuk melatih model *tagger*, karena pada tahap tersebut setiap *n-gram* (penelitian ini terbatas pada 1-gram sampai 2-gram) pada

kedua korpus dihitung probabilitas kemunculannya menggunakan persamaan (2) lalu diberikan peringkat persentil pada kedua gaya teks. *N-gram* yang peringkatnya memenuhi atau melampaui batas yang ditentukanlah yang akan dipakai untuk melatih model *tagger*. Pada penelitian ini, batas yang ditentukan adalah 70% sehingga *n-gram* yang peringkatnya melampaui batasan tersebut akan dipakai untuk melatih model *tagger*. Tabel 1 dan 2 menunjukkan 10 peringkat teratas *unigram*, sedangkan Tabel 3 dan 4 menunjukkan 10 peringkat teratas *bigram* pada korpus sopan dan tidak sopan.

Tabel 1. Kumpulan *unigram* peringkat 10 teratas pada korpus sopan

N-gram	***Percentile Rank (%)***
Mengapa	100
Namun	99.13
Membuat	96.8
Tahu	95.97
Tetapi	95.39
Sepertinya	95.10
Kasihan	94.24
informasinya	92.80
ditindaklanjuti	92.51
Dipakai	91.65

Table 2. Kumpulan *unigram* peringkat 10 teratas pada korpus tidak sopan

N-gram	***Percentile Rank (%)***
kalo	100
udah	99.67
ko	99.34
bener	98.35
aja	98.68
gitu	97.77
rb	97.2
dipake	96.71
gua	96.38
kepotong	96.05

Tabel 3. Kumpulan *bigram* peringkat 10 teratas pada korpus sopan

N-gram	***Percentile Rank (%)***
bagaimana ini	99.71
mengapa saya	99.42
tapi tidak	98.84
seperti itu	98.56
admin mau	98.27
xxxuserxx mengapa	97.98
ini bagaimana	97.69
saja tidak	97.40
tidak dibalas	97.12
yang lalu	96.54

Table 4. Kumpulan *bigram* peringkat 10 teratas pada korpus tidak sopan

N-gram	***Percentile Rank (%)***
ya min	99.01
ini gimana	98.68
xxxnumberxxx rb	97.20
saya ya	95.72
dong min	93.42
min mau	92.77
tapi kok	90.14
kenapa sih	89.15
thank you	88.33
kenapa ya	87.51

Semakin tinggi peringkat *n-gram* pada suatu korpus maka semakin tinggi juga relevansinya terhadap gaya teksnya dan sebaliknya. Sebagai contoh, kata "kalo" pada Tabel 2 memiliki peringkat persentil tertinggi *unigram* pada korpus tidak sopan, hal tersebut menjadikan kata tersebut sangat relevan pada gaya teks tidak sopan namun tidak relevan sama sekali terhadap gaya teks sopan. Jika dibandingkan pada kedua korpus dari *unigram* ataupun *bigram*, kedua gaya teks tersebut memiliki sifat yang berlawanan seperti *bigram* "ini bagaimana" pada korpus sopan yang berlawanan dengan "ini gimana" pada korpus tidak sopan. Dapat dikatakan bahwa polaritas kedua gaya teks pada data sudah cukup baik.

B. Hasil Evaluasi Model Tagger dan Generator

Setelah melakukan pelatihan pada model *Tagger* dan *Generator* dengan iterasi 15 *epoch*, hasil evaluasi dapat dilihat pada Tabel 5.

Tabel 5. Hasil evaluasi model pada masing-masing metrik

BLEU 1	**BLEU 2**	**BLEU 3**	**BLEU 4**	**METEOR**
0.5348	0.403	0.312	0.245	0.262

Hasil evaluasi yang dapat dilihat pada Tabel 5 menunjukkan bahwa model mendapatkan performa terbaik pada **BLEU 1** dan **BLEU 2**, hal tersebut dapat dikatakan teks

yang dihasilkan oleh model mendapatkan rata-rata kecocokan tertinggi pada teks yang terdapat di korpus *test* hanya pada tingkat *unigram* dan *bigram.* Hal tersebut menunjukkan bahwa model belum cukup baik untuk menghasilkan *output* yang sesuai dengan prinsip ***content preservation.***

Untuk pengukuran kefasihan teks yang dihasilkan oleh model atau ***fluency***, dapat dilihat pada contoh di bawah ini:

Input: min pembelian token pln apa ada kendala, ini blm masuk udah xxxnumberxxx jam lebih*?*

Output: admin pembelian token pln apa ada kendala, blm masuk masuk xxxnumberxxx jam lebih?

Hasil yang diharapkan: admin, pembelian token pln apa ada kendala? ini belum masuk sudah xxxnumberxxx jam lebih**.**

Pada contoh di atas, *output* yang dihasilkan oleh model meskipun masih jauh terhadap hasil yang diharapkan namun masih dapat dipahami jika dibaca. Hal tersebut masih membutuhkan peningkatan, karena teks yang dihasilkan masih belum secara utuh berubah menjadi teks yang sopan.

V. KESIMPULAN

Penelitian ini membuat model *Text Style Transfer* untuk mengubah teks tidak sopan menjadi sopan dalam bahasa Indonesia menggunakan strategi mempertahankan bentuk formalitas pada kalimat yang disampaikan. Metode yang digunakan pada penelitian ini adalah *Tag and Generate Approach* yang secara spesifik terdapat dua model yaitu *tagger* dan *generator*, kedua model tersebut memiliki peran masing-masing terhadap teks *input.* Model *tagger* menggantikan *n-gram* yang memiliki karakteristik gaya teks tidak sopan dengan *token tag.* Sedangkan, model *generator* mengisi *token tag* dengan *n-gram* yang memiliki karakteristik gaya teks sopan.

Hasil akhir evaluasi terbaik didapatkan dengan metrik **BLEU-1** dan disusul dengan metrik **BLEU-2** dengan skor masing-masing 0.5348 dan 0.403 dari skala 0 – 1. Hal tersebut dapat diartikan bahwa kesamaan teks yang dihasilkan oleh model dengan *output* yang diharapkan mendapatkan hasil terbaik pada 1 sampai 2 kata. Kefasihan bahasa dari teks yang dihasilkan oleh model cukup baik untuk dipahami, namun gaya teks yang dihasilkan belum secara keseluruhan berubah.

Dapat disimpulkan bahwa model pada penelitian ini dibutuhkan pengembangan lebih untuk dapat diaplikasikan di dunia nyata. Untuk penelitian selanjutnya disarankan untuk menambah jumlah data, menurunkan peringkat persentil *n-gram* bersama ahli bahasa agar polaritas dari kedua gaya teks tersebut dapat terlihat jelas, dan menambah lapisan *transformer* atau mencoba dengan arsitektur LSTM.

REFERENSI

[1] P. Brown, "Politeness and Language," in *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, 2015.

[2] Yosepha Pusparisa, "Tingkat Kesopanan Netizen Indonesia Paling Buruk Se-Asia Pasifik," 2021. https://databoks.katadata.co.id/datapublish/2021/02/26/tingkat-kesopanan-netizen-indonesia-paling-buruk-se-asia-pasifik.

[3] Pranowo, *Berbahasa secara Santun*. Yogyakarta: Pustaka Pelajar, 2012.

[4] Abdurrahman, *Pragmatik: Konsep Dasar Memahami Konteks Tuturan*. 2006.

[5] I. N. Kardana, M. S. Satyawati, and I. G. N. Adi Rajistha, "Strategies to Create Polite Expressions in Indonesian Communication," *Int. J. Linguist.*, vol. 10, no. 6, p. 1, 2018, doi: 10.5296/ijl.v10i6.13851.

[6] A. Madaan *et al.*, "Politeness transfer: A tag and generate approach," *arXiv*. 2020, doi: 10.18653/v1/2020.acl-main.169.

[7] D. Jin, Z. Jin, Z. Hu, O. Vechtomova, and R. Mihalcea, "Deep Learning for Text Style Transfer: A Survey," pp. 1–47, 2020, [Online]. Available: http://arxiv.org/abs/2011.00416.

[8] Z. Hu, R. K.-W. Lee, C. C. Aggarwal, and A. Zhang, "Text Style Transfer: A Review and Experimental Evaluation," 2020, [Online]. Available: http://arxiv.org/abs/2010.12742.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," pp. 311–318, 2002, doi: 10.3115/1073083.1073135.

[10] A. Lavie and A. Agarwal, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *Proc. Second Work. Stat. Mach. Transl.*, no. June, pp. 228–23, 2007, [Online]. Available: http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75.

[11] J. Li, R. Jia, H. He, and P. Liang, "Delete, retrieve, generate: A simple approach to sentiment and style transfer," *NAACL HLT 2018 - 2018 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, pp. 1865–1874, 2018, doi: 10.18653/v1/n18-1169.

[12] C. N. Dos Santos, I. Melnyk, and I. Padhi, "Fighting offensive language on social media with unsupervised text style transfer," *ACL 2018 - 56th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.*, vol. 2, pp. 189–194, 2018, doi: 10.18653/v1/p18-2031.

[13] Y. Zhang, J. Xu, P. Yang, and X. Sun, "Learning sentiment memories for sentiment modification without parallel data," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, pp. 1103–1108, 2020, doi: 10.18653/v1/d18-1138.

[14] R. Kneser and H. Ney, "Improved backing-off for M-gram language modeling," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 1, pp. 181–184, 1995, doi: 10.1109/ICASSP.1995.479394.

[15] H. A. Wibowo *et al.*, "Semi-Supervised Low-Resource Style Transfer of Indonesian Informal to Formal Language with Iterative Forward-Translation," *2020 Int. Conf. Asian Lang. Process. IALP 2020*, pp. 310–315, 2020, doi: 10.1109/IALP51396.2020.9310459.