Laporan Akhir Tugas MK. Data Mining (KOM332), Semester Genap 2019/2020

Pemodelan Topik dan Analisis Sentimen Komentar Warganet Twitter Mengenai COVID-19 Menggunakan Metode *Latent Dirichlet Allocation*

Irfan Alghani Khalid (G64170054), Rafie Fadlurrahman (G64170074), Julius Christoper Siswanto (G64170097), Amin Elhan (G64170109)

Kelompok: 1, Kelas Paralel: K2

Abstrak

Covid-19 menjadi pandemi di dunia. Banyak hal yang terjadi khususnya di Indonesia, sehingga menjadi peluang untuk melihat persepsi masyarakat terhadap suatu kondisi. Analisis sentimen merupakan salah satu cara dengan menggunakan *machine learning* untuk mengetahui persepsi dari *big data.* Twitter merupakan platform *microblogging* yang dapat menampung opini mengenai berbagai topik, temasuk Covid-19. Tahapan yang dilakukan adalah pra-proses, analisis n-gram, analisis metadata, pemodelan topik, lalu dilakukan analisis sentimen. Ditemukan beberapa topik besar yang sering dibicarakan dengan berbagai sentimen. Ada sentimen positif dan negatif yang terlihat, dan beberapa topik yang sering dibicarakan setelah proses LDA. Ditemukan angka 0.25 pada *subjectivity rate* yang berarti sangat sedikit opini yang dapat diolah menggunakan analisis dengan kriteria yang ditentukan. Diperlukan selektif yang lebih dalam memilih data sesuai kebutuhan untuk analisis yang lebih akurat.

Kata Kunci**: covid-19, sentiment analysis, text mining, topic modelling**

PENDAHULUAN

Latar Belakang

Penyebaran COVID-19 yang kian meluas di Indonesia menyebabkan keharusan dalam peningkatan pengetahuan dan kesadaran masyarakat dalam rangka pencegahan dan penanggulangan penyakit. Per tanggal 27 Mei 2020, terdapat 23.851 kasus terkonfirmasi dan 1473 kasus kematian di Indonesia karena COVID-19. Kejadian tersebut merupakan permasalahan yang membutuhkan partisipasi dari seluruh masyarakat dalam penanggulangannya. Namun, di Indonesia masih terdapat banyak kontroversi seputar penyakit ini, baik di kalangan masyarakat maupun tenaga kesehatan, mulai dari aspek manifestasi klinis, penegakan diagnosis, tatalaksana, hingga kebijakan terkait upaya pencegahan dan penanggulangan.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

Melihat pesatnya perkembangan media sosial khususnya *Twitter* di Indonesia, warganet juga meluapkan berbagai macam tanggapan dan informasi yang bisa kita lakukan analisis lebih lanjut lagi dalam mencari pola-pola yang terdapat dalam data tersebut. Melalui metode pemodelan topik dan sentimen analisis, nantinya akan didapatkan topik-topik dominan yang sedang dibicarakan oleh warganet *Twitter* dan melihat berbagai macam emosi positif, negatif maupun netral dalam data tersebut. Hasil tersebut dapat digunakan dalam pengambilan keputusan oleh pemerintah, riset dan penelitian, maupun keperluan lainnya.

Tujuan

Akan dilakukan eksplorasi terhadap data COVID-19 dan dikaitkan dengan kebijakan yang telah diambil oleh pemerintah. Selain itu, ingin diketahui bagaimana pengaruh COVID-19 terhadap opini yang diberikan dan bagaimana sentimen nya apakah positif atau negatif.

Ruang Lingkup

Data yang digunakan diperoleh dari data Covid-19 John Hopkins University dan data pengguna *Twitter* yang memiliki atribut tanggal tweet, user, tweet, jumlah retweet, dan jumlah like dengan jumlah data yang diambil sebanyak 10000 baris data. Teknik yang digunakan yaitu analisis eksplorasi data untuk mencari pola, trend, dan frekuensi kata yang sering muncul (n-gram) dari suatu data, Latent Dirichlet Allocation untuk melakukan topic modelling, serta Sentiment Analysis untuk melakukan penilaian sentimen terhadap tweet.

Manfaat

Penelitian ini diharapkan agar dapat lebih mengetahui relevansi topik, pendapat, dan respon antara kondisi yang terjadi di lapangan dengan yang disampaikan oleh warganet melalui Twitter mengenai pandemik Covid-19 yang nantinya dapat digunakan untuk berbagai macam pengambilan keputusan maupun riset dan penelitian.

TINJAUAN PUSTAKA

Pandemi yang disebabkan oleh virus SARS-CoV-2 ditetapkan sejak tanggal 11 Maret 2020 (World Health Organization, 2020). Banyak dampak sosial dan ekonomi yang terlihat dari keputusan pemerintah yang diambil untuk menangani hal ini. Salah satunya yaitu instruksi presiden untuk bekerja, belajar, dan beribadah di rumah (Sekretariat Kabinet, 2020). Sisi sosial terlihat sebanyak 38 ribu narapidana dibebaskan pada bulan April (Kurniawati, 2020). Pada waktu yang berdekatan, terlihat dari sisi ekonomi bahwa jumlah korban PHK menurut APINDO mencapai tujuh juta orang pada bulan April (Wardoyo, 2020). Pemerintah melakukan kebijakan seperti membuat Gugus

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

Tugas Percepatan Penangan COVID-19, pemerintah daerah melakukan mengajukan PSBB ke Menteri Kesehatan dan Bansos di DKI berupa BLT dan bahan makanan pokok mulai disalurkan. Tentunya keputusan yang dibuat tidak lepas dari reaksi warga yang merasakan dampaknya secara langsung maupun tidak langsung. Menurut Huang et al. (2009) twitter merupakan *microblogging* yang populer untuk mengekspresikan opini terhadap berbagai topik.

Menurut Torfi et al. (2020), Pemrosesan bahasa alami atau *Natural Language Processing (NLP)* merupakan salah satu bidang yang merupakan gabungan dari Ilmu Komputer dan Ilmu Linguistik. Bidang ini memanfaatkan kemampuan komputasi untuk melakukan ekstraksi informasi yang tersembunyi dari data. Contoh pemrosesan yang termasuk ke dalam NLP, yakni analisis sentimen, pemodelan topik, ekstraksi entitas, dan lain-lain. Pemodelan topik merupakan salah satu bagian dari NLP yang digunakan untuk mengekstraksi topik yang terdapat pada teks. Teks tersebut dapat dikelompokkan ke beberapa topik dengan pendekatan probabilistik. Algoritme yang menggunakan pendekatan probabilistik tersebut bernama *Latent Dirichlet Allocation (LDA)* yang digagas oleh David Blei pada tahun 2003. Analisis sentimen merupakan salah satu bagian dari NLP yang mengekstraksi informasi yang bersifat subjektif. Yang dimaksud dengan subjektif adalah teks yang bersifat opini. Selain itu, teks tersebut akan diperiksa bagaimana polaritasnya terhadap suatu bahasan apakah bersifat positif atau negatif.

METODE

Data

Data yang digunakan diperoleh dari data Covid-19 John Hopkins University dan data pengguna Twitter yang memiliki atribut tanggal tweet, user, tweet, jumlah retweet, dan jumlah like dengan jumlah data yang diambil sebanyak 10000 baris data dalam format csv.

Tahapan Kegiatan

- Pengambilan data
    - Dataset adalah Data Twitter yang diambil menggunakan package “rtweet” dengan keyword “virus”, “corona”, “covid-19”, dan “covid19” dan parameter bahasa = “id”.
- Pra Proses Data
    - Melakukan penghapusan *mention, hashtag, link,* angka, dan stopword, dan setelah itu diterjemahkan untuk mempermudah analisa.
- Eksplorasi Data
    - Melakukan analisis terhadap data dalam bentuk *n-gram*, *wordcloud*, dan analisis sentimen berdasarkan polaritas dan subjektivitas.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

- Pemodelan Data
    - Metode *Latent Dirichlet Allocation* digunakan dalam pemodelan data untuk mencari topik-topik yang terdapat dalam data.

Lingkungan Pengembangan

Spesifikasi perangkat keras yang digunakan adalah Intel I7-7700HQ untuk CPU, 8GB RAM, dan NVIDIA GTX 1050 Ti untuk GPU. Perangkat lunak yang diperlukan yaitu RStudio untuk melakukan *programming* dalam bahasa R dan Jupiter Notebook untuk bahasa Python. Perangkat keras yang digunakan yaitu sebuah laptop.

HASIL DAN PEMBAHASAN

Pengambilan Data

Pengambilan Dataset yang diambil untuk project ini adalah Data Twitter yang diambil menggunakan package "rtweet" sejumlah 100 ribu. Data yang diambil bersifat acak dan data yang diambil di-tweet pada tanggal 26 April 2020 sampai 27 April 2019 dengan keyword "virus", "corona", "covid-19", dan "covid19" dan parameter bahasa = "id". Dengan catatan bahwa terdapat tweet yang tidak dalam bahasa Indonesia, misal Melayu, Hindi, dan Urdu. Data disampling menjadi 10000.

Pra Proses Data

Setelah dilakukan proses pengambilan data, maka dilakukanlah pra-proses terhadap data tweet tersebut. Pada data tersebut, hal pertama yang perlu dilakukan adalah membuat semua kata di tweet dalam bentuk *lowercase.* Lalu, hal yang perlu dihilangkan adalah *mention, hashtag, link,* angka, dan stopword. *Stopword* ini adalah kata yang tidak memiliki pengaruh terhadap pemaknaan suatu kalimat. Kata-kata yang termasuk *stopword* misalnya adalah saya, anda, yang, bagaimana, dan lain-lain. Lalu, tweet tersebut diterjemahkan dalam bahasa inggris. Untuk penerjemahan, kami menggunakan *Google Translate* API. Alasan perlu dilakukan penerjemahan adalah untuk memudahkan pemrosesan data, lalu kebanyakan library untuk pemrosesan bahasa dapat bekerja maksimal untuk tweet berbahasa Inggris, lalu karena pemrosesan kata-kata *slang* dan yang disingkat, seperti yg, sy, km, dll, dapat diperbaiki dengan menerjemahkan ke bahasa inggris, dan terakhir karena tweet yang didapatkan tidak semuanya berbahasa Indonesia sehingga perlu diterjemahkan.

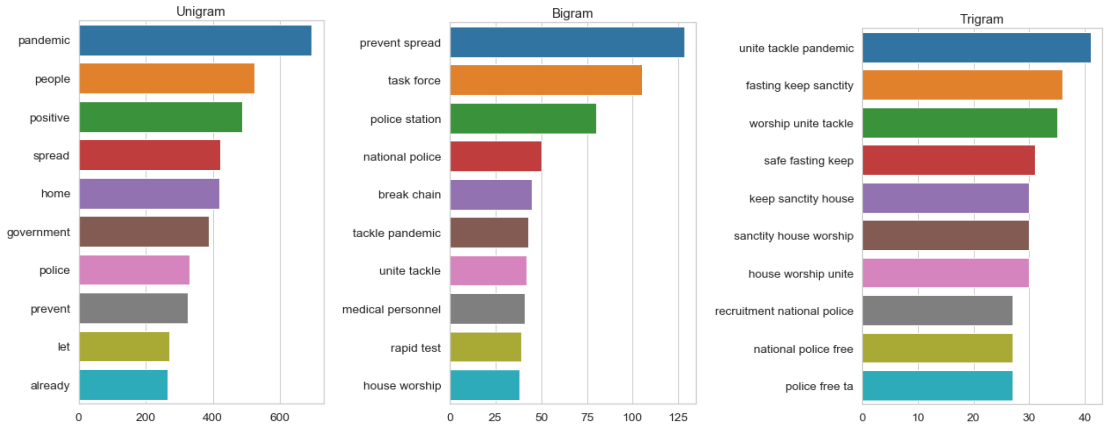
Setelah diterjemahkan, maka dapat dilakukan proses analisis dan pemodelan. Untuk mendapatkan hasil analisis yang baik, maka keyword yang digunakan untuk pengambilan tweet dihapus. Alasannya adalah karena kata tersebut pasti akan menjadi kata yang sering muncul dimana hal yang ingin diperhatikan adalah kata-kata di luar keyword tersebut.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

Analisis N-Gram

Sebelum melakukan pemodelan, akan dilakukan analisis *n-gram* terhadap *dataset* tersebut. Analisis *n-gram* merupakan salah satu analisis di *text mining* yang digunakan untuk mengetahui kata atau kalimat apa yang sering muncul di *dataset* tersebut. Huruf n yang dimaksud dari *n-gram* adalah jumlah kata yang dianalisis. 1-gram untuk mencari frekuensi satu kata, 2-gram untuk mencari frekuensi 2 kata gabungan yang sering muncul, dan seterusnya.

Untuk hasil analisis dari n-gram dapat dilihat pada grafik di bawah,

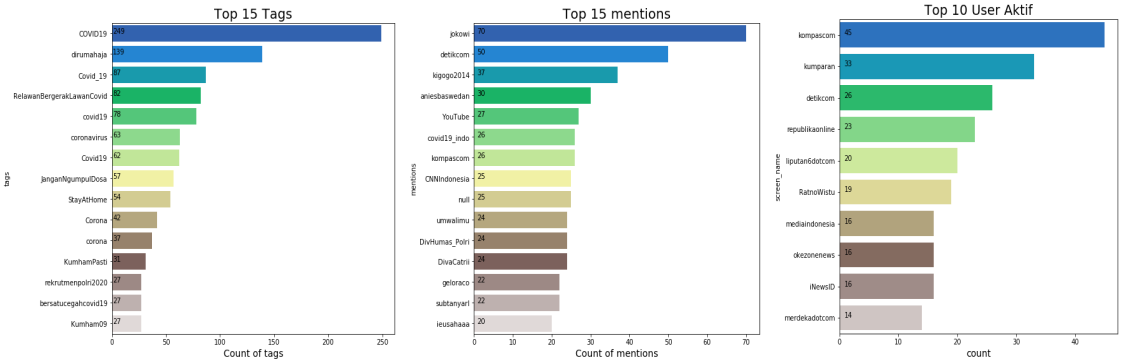


Berdasarkan grafik diatas, pada grafik 1-gram, terlihat kata yang sering muncul adalah kata berkaitan dengan COVID-19, seperti pandemi, orang, positif, mencegah, rumah, penyebaran, dan sebagainya. Lalu, pada grafik 2-gram, terlihat bahwa kalimat yang sering muncul adalah gugus tugas, kantor polisi, hilangkan rantai, lawan pandemi, bersatu melawan, ibadah di rumah, dan sebagainya. Terakhir, pada grafik 3-gram, terlihat bahwa kalimat yang sering muncul adalah bersatu melawan pandemi, beribadah bersatu melawan, ibadah di rumah bersatu, dan sebagainya. Secara umum, grafik tersebut terdapat kata dan kalimat yang berkaitan dengan virus corona, rekrutmen polri, dan beribadah di rumah karena pengambilan tweet bertepatan dengan bulan puasa.

Analisis Metadata

Selain itu, dilakukan juga analisis jumlah mention, hashtag, dan user aktif. Berikut grafik yang dihasilkan,

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.



Berdasarkan grafik di atas, terlihat bahwa *hashtag* terpopuler berkaitan dengan COVID-19. Selain itu, terdapat hashtag seperti JanganNgumpulDosa, KumhamPasti, dan rekrutmenpolri2020. Hal ini sering muncul di twitter, khususnya Indonesia, untuk memudahkan marketing pihak tertentu. Sehingga, hal ini menjadi tantangan dalam mengolah tweet bahasa Indonesia yang di dalamnya terdapat tweet di luar konteks bahasan.

Akun jokowi, detikcom, dan aniesbaswedan mendapatkan mention terbanyak untuk tweet berbahasa Indonesia. Selain itu, terdapat akun yang memiliki mention terbanyak, yakni kigogo2014 yang merupakan akun yang berasal dari wilayah Afrika. Untuk user teraktif, semua akun yang muncul di grafik tersebut adalah akun media massa. Contohnya adalah kompascom, kumparan, detikcom, republikaonline, liputan6com, dan sebagainya.

Wordcloud

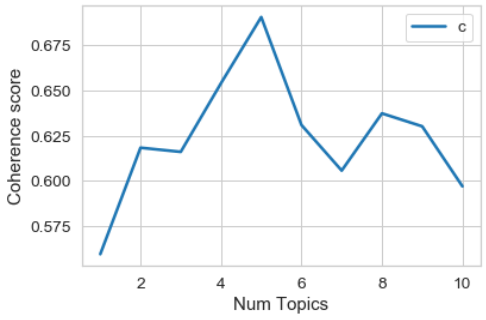
Analisis eksplorasi terakhir yang dapat dilakukan pada *text mining* adalah visualisasi kata tersebut ke dalam bentuk *wordcloud. Wordcloud* merupakan sebuah visualisasi dari kumpulan kata yang bergerombol dengan ukuran yang berbeda. Hal tersebut didapatkan berdasarkan jumlah kata yang ada pada *dataset.* Untuk *wordcloud* yang dihasilkan berbentuk seperti berikut,

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

Pemodelan Topik

Setelah melakukan analisis eksplorasi, hal yang perlu dilakukan selanjutnya adalah melakukan pemodelan topik terhadap dataset. Untuk pemodelan topik, digunakan algoritma *unsupervised Latent Dirichlet Allocation (LDA)* dimana algoritme ini akan mengelompokkan *tweet* ke dalam topik dengan pendekatan probabilistik. Untuk implementasi LDA akan digunakan library gensim dengan visualisasi menggunakan LDAvis.

Untuk mendapatkan performa model yang maksimal, maka dilakukan pemilihan jumlah topik yang dapat mencakup keseluruhan dataset dan menggambarkan topik tersebut berdasarkan skor koheren.




```
Num Topics = 1  has Coherence Value of 0.559663
Num Topics = 2  has Coherence Value of 0.6184
Num Topics = 3  has Coherence Value of 0.61612
Num Topics = 4  has Coherence Value of 0.654003
Num Topics = 5  has Coherence Value of 0.690515
Num Topics = 6  has Coherence Value of 0.630942
Num Topics = 7  has Coherence Value of 0.605789
Num Topics = 8  has Coherence Value of 0.637356
Num Topics = 9  has Coherence Value of 0.630217
Num Topics = 10  has Coherence Value of 0.597111
```

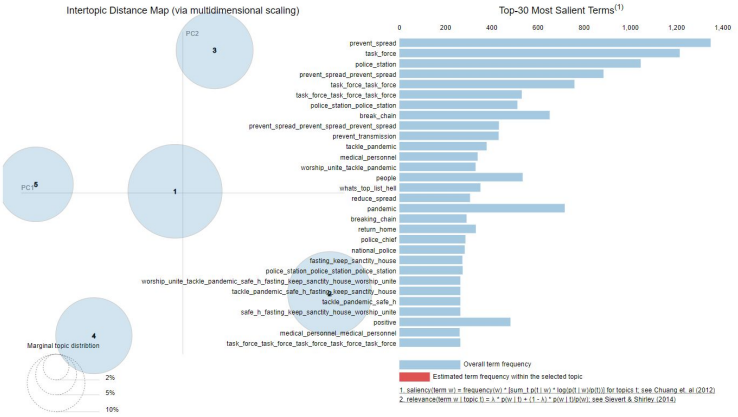
Berdasarkan plot skor koheren (tingkat kesamaan semantik), akan dipilih sebanyak 5 topik yang akan menjadi acuan dalam membuat model dengan nilai skor koheren sebesar 0.690515. Setelah menentukan jumlah topik yang tepat untuk dataset, maka dilakukan proses pemodelan menggunakan LDA.

```
Topics found via LDA:

Topic #0:
people home china god fear miss work chain die news

Topic #1:
police spread prevent pandemic government village force task update week

Topic #2:
positive fasting pandemic house news home money country government worship

Topic #3:
pandemic medical government social handling minister new patient hospital affected

Topic #4:
finished positive let want people world pandemic plague conspiracy meet
```

Berdasarkan hasil di atas, telah didapatkan 5 topik. Tema topik dapat dilihat dari besarnya peluang kata yang sering muncul pada tweet yang termasuk ke dalam

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

kelompok tersebut. Topik 0 membahas dampak negatif yang diakibatkan oleh COVID-19. Topik 1 membahas terkait dengan gugus tugas, polisi, dan instansi lain dalam mencegah penyebaran COVID-19. Topik 2 berkaitan dengan puasa dan ibadah yang terpengaruh oleh COVID-19. Topik 3 berkaitan dengan pemerintah dan kementrian kesehatan. Topik 4 berkaitan dengan teori konspirasi dan berakhirnya COVID-19.

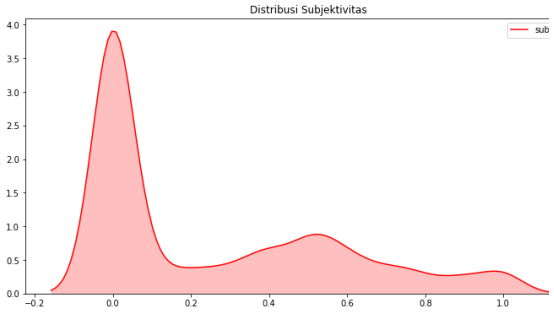


Visualisasi dengan menggunakan LDAvis menunjukkan bahwa kelima model tersebut tidak saling bersinggungan dan memiliki *cluster* masing-masing. Kata yang paling menonjol dari keseluruhan topik adalah *prevent spread, task force* dan *police station*. Ukuran dari tiap *cluster* menunjukkan seberapa penting topik tersebut relatif terhadap data. Topik dengan ukuran terbesar yaitu topik yang membahas tentang tenaga medis, Kementerian Kesehatan, dan pemerintahan.

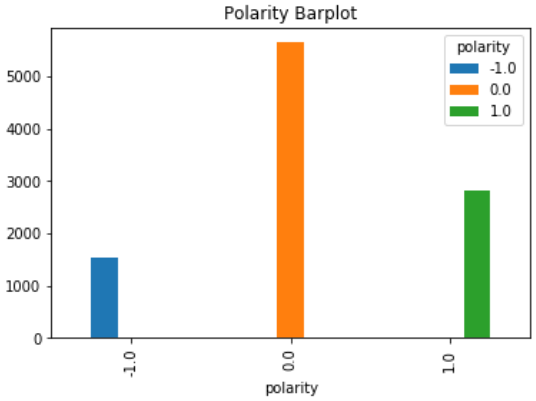
Analisis Sentimen

Analisis sentimen, atau istilah lainnya opinion mining, merupakan analisis yang dilakukan untuk mengetahui bagaimana opini seseorang terhadap suatu topik (subjektivitas) dan bagaimana sikap yang diberikan, apakah positif atau negatif (polaritas). Berikut grafik beserta nilai dari subjektivitas dan polaritas dari dataset yang digunakan.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.




```
Subjectivity Result:

Mean:  0.25368106177684735
```



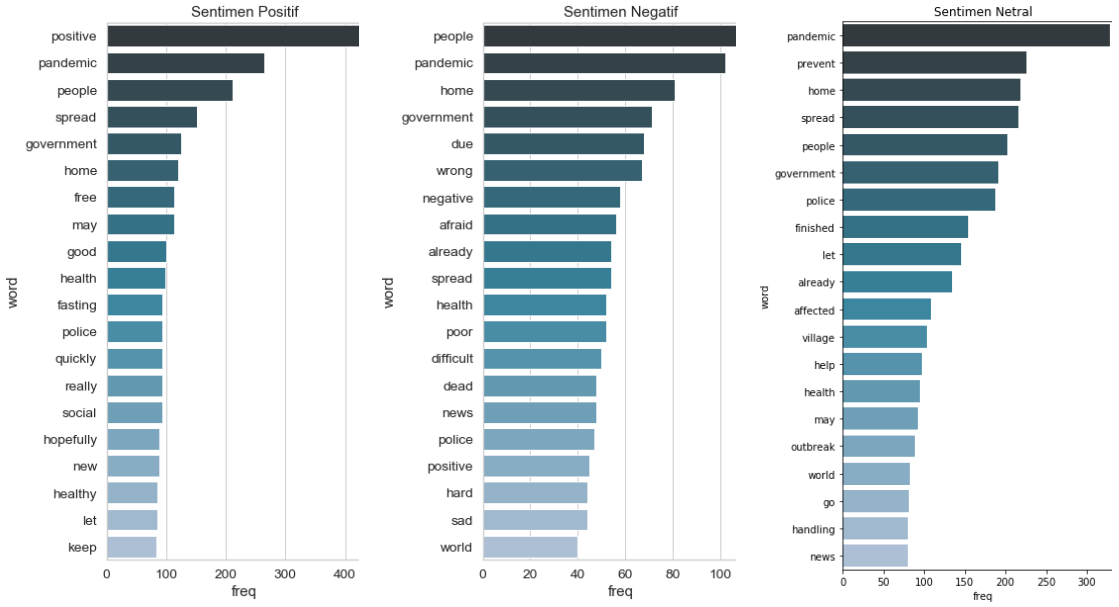

```
Polarity Result:

Positive:  2823
Neutral:  5652
Negative:  1525
Mean:  0.037655186322606116
```

Berdasarkan hasil yang didapatkan, terlihat bahwa nilai tengah dari subjektivitas sebesar 0,254. Hasil ini menandakan bahwa mayoritas tweet pada dataset tersebut bersifat objektif. Hal ini juga terlihat pada mayoritas tweet yang didapatkan adalah berupa berita atau informasi yang menyajikan data-data yang valid. Faktor yang menyebabkan didapatkan hasil yang subjektif karena pengambilan data dari search API dimana dengan menggunakan tweet tersebut tentu tidak hanya tweet berupa opini yang diambil, tweet yang berisi informasi seperti berita juga diambil.

Berikutnya, dari polaritasnya, didapatkan bahwa sentiment yang dihasilkan mayoritas netral atau 0 dengan jumlah tweet sebanyak 5652. Lalu, tweet positif atau +1 sebanyak 2823. Terakhir, tweet negatif atau -1 sebanyak 1525. Berdasarkan nilai tersebut, sangat wajar bahwa sentimen netral menjadi tweet terbanyak, ditambah lagi dengan nilai subjektivitas yang rendah sehingga mendukung pendapat bahwa tweet tersebut mayoritas bersentimen netral. Uniknya lagi, meskipun COVID-19 merupakan suatu hal yang negatif, tweet yang bernilai negatif sangat sedikit dibanding netral dan positif. Dengan ini, disimpulkan bahwa dataset tersebut memiliki banyak tweet yang objektif dan juga banyak tweet yang bernilai positif dibandingkan dengan negatif.

Untuk kata yang sering muncul pada setiap kategori sentimen, berikut grafik yang menampilkan kata yang sering muncul yang dibagi berdasarkan nilai sentimennya,

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.



Berdasarkan plot frekuensi kemunculan kata terbanyak, *pandemic* merupakan kata yang paling sering muncul dari ketiga kategori analisis sentimen. Untuk setiap kategorinya, pada sentimen positif, terlihat banyak kata yang memiliki nada positif, seperti *new, free, good, hopefully,* dan sebagainya. Untuk sentimen negatif, banyak kata yang memiliki makna buruk, seperti *hard, poor, difficult, sad,* dan lain-lain. Untuk netral, terlihat kata-kata yang cenderung bersifat objektif dan tidak ada yang menyatakan perasaan positif maupun negatif.

KESIMPULAN DAN SARAN

Berdasarkan hasil analisis tersebut, maka dapat disimpulkan bahwa COVID-19 benar-benar memberikan pengaruh terhadap informasi yang telah menyebar di Twitter, mulai dari berita mengenai COVID-19, harapan dan keresahan masyarakat, konspirasi berkaitan dengan COVID-19, bahkan ada tweet yang berisikan promosi seperti rekrutmen polri. Hal ini sangat wajar ditemukan di *tweet* orang Indonesia yang memanfaatkan keyword trending dalam menjual usaha mereka. Sayangnya, hasil dari analisis sentimen bisa dibilang kurang baik karena banyak sekali tweet yang bernilai netral dan juga nilai subjektivitas pada dataset tersebut di bawah 0.5 sehingga mayoritas tweet berisikan informasi, bukan opini. Diperlukan juga untuk mengambil kategori yang pas dalam mengambil data sehingga *objectivity rate* yang diperlukan untuk analisis sentimen yang lebih akurat. Semisal mengambil *tweet* dengan *mention* ke akun berita online.

Oleh karena itu, terdapat beberapa catatan penting dari hasil projek ini seperti masih ada tweet yang bukan bahasa Indonesia masuk ke dalam dataset, stopwords yang

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

masih muncul seperti “yg, ga, kalo, udah, dll”. Untuk waktu yang didapatkan masih kurang, mengingat Twitter API yang digunakan versi gratis. Selanjutnya perlu dipastikan dataset hanya memiliki satu jenis bahasa dan terdapat *library* yang dapat digunakan untuk mengolah data dalam bahasa tersebut. Menggunakan *hyperparameter tuning* dalam model untuk mendapatkan model dengan parameter terbaik perlu dipertimbangkan. Penambahan stopword baru (*slang*) diperlukan untuk meningkatkan akurasi dari hasil. Dibutuh Twitter API premium sehingga analisis yang dilakukan bisa lebih baik dan informatif. Terakhir, perlu diperhatikan metode apa yang dilakukan untuk pengambilan topik. Hal ini berpengaruh terhadap analisis sentimen. Karena analisis sentimen bersifat opini, maka sebaiknya tidak mengambil dataset menggunakan search API karena dipastikan *tweet* yang didapatkan bersifat campuran antara fakta dan opini. Untuk mencegah hal tersebut, pengambilan data dapat dilakukan dengan mengambil mention masyarakat terhadap suatu akun berita. Dengan begitu, analisis sentimen dapat menghasilkan *insight* yang lebih baik.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680
*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.

DAFTAR PUSTAKA

Go, A., Huang, L. and Bhayani, R., 2009. Twitter sentiment analysis. *Entropy*, *17*, P.252.

Kaur, Chhinder & Sharma, Anand. 2020. *Twitter Sentiment Analysis on Coronavirus using Textblob* [Internet]. [Diunduh pada 28 Mei 2020]. Tersedia pada: https://www.researchgate.net/publication/339998775_Twitter_Sentiment_Analysis_on_Coronavirus_using_Textblob.

Severyn, A. and Moschitti, A., 2015, August. Twitter sentiment analysis with deep Convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 959-962).

Torfi, Amirsina & Shirvani, Rouzbeh & Keneshloo, Yaser & Tavvaf, Nader & Fox, Edward, 2020. Natural Language Processing Advancements By Deep Learning: A Survey.

Wang, H., Can, D., Kazemzadeh, A., Bar, F. and Narayanan, S., 2012, July. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the ACL 2012 system demonstrations* (pp. 115-120). Association for Computational Linguistics.

[WHO] World Health Organization. 2020. WHO Director-General's opening remarks at the media briefing on COVID-19 [Internet]. [Diunduh pada 28 Mei 2020]. Tersedia pada: https://www.who.int/dg/speeches/detail/who-director-generals-opening-remarks-at-the-media-briefing-on-covid-19---11- march-2020.

Zhang, L., Ghosh, R., Dekhil, M., Hsu, M. and Liu, B., 2011. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *HP Laboratories, TechnicalReport HPL-2011*, *89*.

Departemen Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, Bogor 16680

*Mahasiswa Program Studi Ilmu Komputer, FMIPA-IPB; Surel: amin_elhan@apps.ipb.ac.id, julius_chris@apps.ipb.ac.id, irfan_khalid@apps.ipb.ac.id, rafie_fadlurahman@apps.ipb.ac.id.