

Sentiment Classification untuk Opini Berita Sepak Bola

Eka Rahayu Setyaningsih, Program Studi Informatika Institut Sains dan Teknologi Terpadu Surabaya

Abstrak— Pada penelitian ini akan dibahas mengenai sebuah aplikasi yang dibuat secara khusus untuk mengkategorikan opini masyarakat terhadap sebuah berita Sepak Bola. Opini yang diolah diperoleh dari dua sumber, yaitu melalui hasil crawl situs berita olah raga dan opini yang ditambahkan oleh user sendiri pada aplikasi ini. Opini yang ada nantinya akan disajikan secara terpisah menurut kelompoknya; sentiment positive, negative, maupun netral. Proses klasifikasinya sendiri terdiri dari dua tahap. Tahap pertama adalah proses preprocessing yang terdiri atas proses tokenisasi, normalisasi, case folding, stop word removing, common word removing, stemming. Tahap kedua adalah mengklasifikasikan opini-opini tersebut dengan algoritma Baseline, dan Naive Bayes. Opini yang digunakan untuk proses klasifikasi yaitu opini yang menggunakan bahasa Inggris dari situs fifa.com dan goal.com. Dari perhitungan macroaveraged untuk setiap kelas, didapatkan akurasi 93,06%, presisi 81,90%, dan recall 92,67% untuk kelas sentiment positive. Dari perhitungan kelas sentiment negative didapatkan akurasi 87,73%, presisi 96,29%, dan recall 83,63%. Dari perhitungan kelas sentiment netral didapatkan akurasi 92,26%, presisi 64,44%, dan recall 90,37%. Kesimpulan yang diperoleh saat penelitian ini dari awal hingga akhir adalah, proses crawling yang digunakan untuk mendapatkan berita dan komentar berita sangat membantu dalam penambahan konten website, tetapi banyak sekali komentar berita yang diperoleh kurang cocok untuk proses klasifikasi.

Kata Kunci—Sentiment Analysis, Opinion Classification, Naive Bayes, Football.

I. PENDAHULUAN

Informasi yang tersedia dalam dunia web dibagi menjadi dua tipe yaitu fakta dan opini. Setiap orang bebas untuk mengekspresikan opininya dalam berbagai ragam topik dan melalui berbagai macam media online seperti blog, jejaring sosial dan lain-lain. Salah satu bentuk penyajian informasi tekstual yaitu Sentiment Analysis.

Belakangan, sentiment analysis menarik sebagian besar perhatian baik dari akademis maupun industri [1]–[5]. Hal ini disebabkan karena banyaknya masalah penelitian dan berbagai macam aplikasi yang menantang. Dengan menggunakan sentiment analysis, seseorang dapat melihat tanggapan seseorang terhadap sesuatu permasalahan yang diamati sebelum menyimpulkan sebuah keputusan.

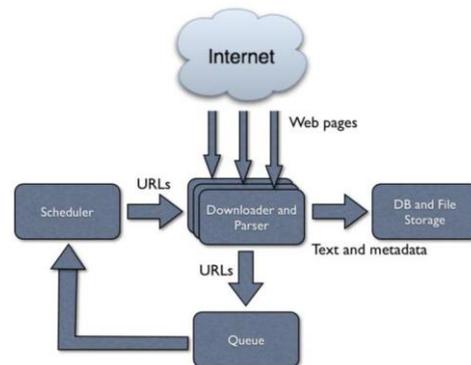
Berita sepak bola adalah salah satu topik yang sering digunakan oleh seseorang untuk memberikan opini. Opini yang diberikan bisa sebagai opini positif atau opini negatif maupun opini netral. Untuk dapat mengetahui opini tersebut

sebagai opini positif atau negatif atau netral, pengguna harus membaca dan mengamati opini tersebut satu per satu. Oleh karena itu, untuk mempermudah pengguna mengetahui opini tersebut masuk ke opini positif atau negatif atau netral.

II. TINJAUAN PUSTAKA

A. Web Crawling

Web Crawling merupakan suatu proses yang digunakan untuk menjelajah dan mengambil sekumpulan halaman dari sebuah web. Untuk melakukan proses tersebut dibutuhkan sebuah perangkat lunak yang disebut web crawler. Web crawler adalah salah satu komponen penting dalam mesin pencari modern seperti Google, Yahoo, dan lain-lain. Web crawler sering dikenal dengan nama web spider atau web robot.



Gambar 1. Cara Kerja Crawler

Seperti yang ditunjukkan pada gambar 1, mula-mula web crawler akan memulai kerjanya dengan mengunjungi situs yang sudah disebutkan oleh user sebagai *url seed* atau URL yang ada di dalam sebuah database yang digunakan untuk menyimpan alamat situs yang ingin dikunjungi. Database penyimpanan alamat situs tersebut dikenal dengan sebutan frontier url.

Setelah web crawler tiba atau sampai pada halaman website yang ditentukan, maka web crawler akan melakukan proses fetching. Proses fetching adalah proses yang dilakukan oleh web crawler untuk mengambil dokumen HTML yang terdapat dalam suatu halaman website. Hasil fetching web crawler yang berupa data dan meta data akan disimpan ke dalam penyimpanan utama. Sedangkan hasil penguraian web crawler berupa outlink atau tautan lain yang didapat, akan dimasukkan ke dalam queue.

Bersamaan dengan berjalannya proses crawling, seluruh

outlink yang didapat dari sebuah halaman website akan didaftarkan dalam sebuah queue yang berupa daftar outlink yang ada. Setelah proses fetching pada halaman tersebut selesai, web crawler akan melakukan proses fetching kembali. Selanjutnya, proses fetching akan dilakukan pada halaman web lain yang telah terdaftar dalam queue tersebut. Apabila dalam proses fetching selanjutnya terdapat outlink yang telah terdapat dalam queue, maka outlink tersebut tidak akan diproses dan ditambahkan ke dalam queue. Proses ini akan dilakukan web crawler sampai outlink yang terdaftar dalam queue habis atau batasan kedalaman crawling yang ditentukan oleh user.

B. Sentiment Classification

Sentiment classification merupakan suatu proses mengklasifikasikan teks yang terdapat pada suatu bacaan atau dokumen. Sentiment classification adalah sebuah teknik pengklasifikasian yang mana pembelajarannya menggunakan metode supervised learning. Target output dari metode ini biasanya diambil dari suatu dataset yang telah ada.

Sentiment classification bertujuan untuk menentukan tanggapan dan sikap seseorang sehubungan dengan beberapa topik [6], [7]. Sering kali, sentiment classification digunakan untuk menentukan apakah penulis menyukai atau tidak menyukai sebuah produk dari review yang mereka tulis. Input dari sentiment classifier selalu sebuah teks beropini. Teks beropini adalah teks yang mengandung sentimen positif, atau negatif. Sedangkan output dari sentiment classifier yaitu penggolongan teks ke sebuah kelas seperti positif atau negatif.

Untuk menyelesaikan permasalahan dari Sentiment Analysis, ada banyak algoritma yang dapat digunakan seperti algoritma Naive Bayes, Support Vector Machine, C45, K-Nearest Neighbor, K-Means, Maximum Entropy dan lain-lain. Algoritma yang paling sering digunakan yaitu Naive Bayes dan Support Vector Machine [8]. Naive Bayes sering digunakan karena sederhana tetapi memiliki akurasi yang tinggi sedangkan Support Vector Machine dikarenakan pengerjaan yang sangat baik pada data dengan banyak dimensi. Dalam penelitian ini algoritma yang digunakan yaitu algoritma Naive Bayes dan algoritma Baseline.

C. Baseline Algorithm

Algoritma Baseline [9] melakukan analisis sentimen pada komentar-komentar dengan menggunakan daftar kata positif dan negatif. Algoritma ini akan membandingkan setiap kalimat yang terdapat satu komentar. Kata-kata yang terdapat pada setiap kalimat hanya dibandingkan dengan daftar yang ada dan dikelompokkan langsung tanpa diproses terlebih dahulu kebenarannya kalimat tersebut positif atau negatif.

Segmen Program 1- Algoritma Baseline

```

1. L1={List of Positif words}
2. L2={List of Negatif words}
3. for each commentc in database
   do
4.   for each words w in c do
5.     increment count of positif words
       in L1 that contained in w
6.     increment count of negatif words in L2
       that contained in w
7.   end
8.   if count of negatif word > count of positif

```

```

word
9.   return "negatif"
10.  else return "positif"
11. end

```

Metode Baseline yang akan digunakan dalam penelitian ini adalah supervised baseline. Supervised baseline adalah metode baseline yang mana daftar kata-kata positif dan negatif yang digunakan untuk perbandingan sudah tersedia dan tidak diperlukannya pembelajaran secara otomatis. Daftar kata positif yang digunakan seperti adorable, capable, cool, fair, dan lain-lain. Daftar kata negatif yang digunakan seperti arrogant, awful, selfish, rough, careless, dan lain-lain.

Berikut adalah contoh komentar yang akan dibandingkan dengan daftar kata-kata positif dan negatif yang ada.

Contoh komentar:

"I agree with him on most things, but Ferguson was a great coach, The coach has to have a bigger personality than the players at a big club and he definitely was, That's why they were so successful. Moving on to england now, he's absolutely right, The media and pundits overhype every little talent".

Hasil analisa komentar:

- Kata Positif : 5 buah
- Kata Negatif : 1 buah

Output: Komentar 1 termasuk komentar positif.

TABEL I
DAFTAR KATA POSITIF DAN NEGATIF

Daftar Kata Positif	Daftar Kata Negatif
Abound	abnorm
Absolut	abolish
Absorb	abomin
Abstemiou	abort
Abund	abrad
Abundance	abras
Accept	abrupt

Dengan terlebih dahulu membentuk daftar kata negative dan daftar kata positif yang akan digunakan selama penelitian (seperti yang ditunjukkan pada table 1), maka dapat disimpulkan bahwa dari komentar pertama dapat diketahui bahwa komentar tersebut termasuk komentar positif karena jumlah kata positif lebih banyak dari jumlah kata negatif. Kata-kata yang termasuk kata positif yaitu agree, great, bigger, successfull, dan right. Kata-kata yang termasuk kata negatif yaitu overhype.

D. Naive Bayes Algorithm

Naive Bayes Classifier (NBC) merupakan metode pembelajaran dengan konsep probabilitas sederhana. NBC menggunakan teorema kuno, warisan abad ke-18, yang ditemukan oleh Thomas Bayes. NBC menyertakan dokumen klasifikasi terbimbing, metode pembelajaran yang menghasilkan fungsi untuk memetakan masukan ke keluaran yang diinginkan. NBC menganggap kemunculan satu kata tidak mempengaruhi kemunculan kata lainnya. NBC mampu memberikan kinerja yang cukup baik untuk banyak kasus

modern dengan data yang besar. Adapun untuk menghitung probabilitas fitur kata menggunakan persamaan (1):

$$\hat{P}(w|c) = \frac{\text{count}(w,c)+1}{\text{count}(c)+|V|} \dots\dots\dots (1)$$

Kemudian untuk menghitung probabilitas prior menggunakan persamaan (2):

$$\hat{P}(c) = \frac{N_c}{N} \dots\dots\dots (2)$$

Terakhir untuk menentukan sentimen menggunakan persamaan (3):

$$c = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(w_i|c) \dots\dots\dots (3)$$

E. Language Identification

Language identification merupakan suatu proses identifikasi bahasa. Proses identifikasi bahasa ini bertujuan untuk mengetahui sebuah teks yang tertulis termasuk ke dalam golongan bahasa tertentu. Teks yang teridentifikasi bisa termasuk bahasa Inggris, Indonesia atau bahasa-bahasa lainnya. Proses pengidentifikasian bahasa ini, akan mencocokkan setiap kata yang terdapat pada suatu teks dengan database yang ada. Berikut adalah contoh pengidentifikasian sebuah teks. Contoh teks yang akan diidentifikasi yaitu "worst player ever". Setelah diidentifikasi, teks ini digolongkan dalam bahasa Inggris. Setiap pengidentifikasian akan memiliki nilai angka yang menunjukkan ketepatan pengidentifikasian bahasa. Jika suatu teks digolongkan ke dalam 2 bahasa, maka nilai pengidentifikasian bahasa yang tertinggi akan menjadi penentu bahasa.

Penggunaan Language Detection API dapat dilakukan dengan 2 cara. Cara pertama yaitu menggunakan API yang telah disediakan untuk beberapa bahasa pemrograman seperti Ruby, Java, Python, dan PHP. Karena bahasa pemrograman yang digunakan dalam penelitian ini tidak terdaftar, maka akan menggunakan cara yang lain. Cara yang lain yaitu dengan metode GET dan POST request. Metode POST lebih disarankan untuk jumlah request yang lebih banyak. Kedua metode ini akan memberikan hasil dalam format json. Cara penggunaan GET dan POST request ini cukup mudah, dengan mengakses ke alamat url <http://ws.detectlanguage.com/0.2/detect?parameter>. Ada dua parameter yang digunakan yaitu q dan key. Berikut penjelasan parameter yang digunakan:

1. Parameter q digunakan untuk memasukan sebuah kata atau text yang akan di deteksi. Tidak ada batasan karakter untuk parameter ini.
2. Parameter key ini akan diisi dengan API key yang didapat setelah melakukan pendaftaran pada detectlanguage.com

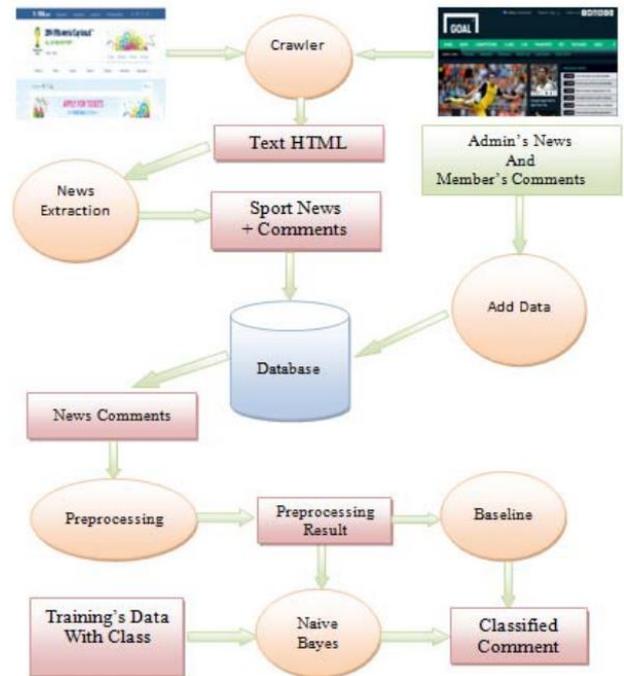
Berdasarkan kedua parameter tersebut untuk menggunakan GET request dapat dilihat pada URL berikut: "http://ws.detectlanguage.com/0.2/detect?q=buenos+dias+senor&key=demo". Dari contoh tersebut, berarti teks yang diinginkan dideteksi adalah buenos dias senor dan key yang digunakan adalah demo. Jika URL itu dijalankan pada

web browser maka akan menampilkan sebuah halaman yang berisi hasil pendeteksian bahasa. Setiap hasil pendeteksian akan berisi kode dari bahasa yang digunakan, nilai keakuratan bahasa, dan status bahasa yang terdeteksi dapat dipercaya atau tidak. Language Detection API untuk sementara mendeteksi sebanyak 83 bahasa.

Untuk mendeteksi beberapa teks, dapat dilakukan dalam satu URL. Berikut adalah contoh pendeteksian: [http://ws.detectlanguage.com/0.2/detect?q \[\]=buenos+dias &q\[\]=morning&key=demo](http://ws.detectlanguage.com/0.2/detect?q[]=buenos+dias&q[]=morning&key=demo). Dalam pendeteksian tersebut, setiap parameter q yang digunakan akan dihitung satu parameter. Jadi untuk pendeteksian tersebut, terdapat dua request yang dilakukan. Hasil request yang dilakukan akan sesuai dengan urutan parameter q yang dikirimkan. Dengan melakukan banyak request dalam satu URL menghemat kuota jaringan dikarenakan jumlah kuota yang terbatas. Untuk penelitian ini digunakan Free Plans untuk language detection API yang terbatas untuk 5000 request dan 1MB kuota perharinya.

III. ARSITEKTUR SISTEM

Gambaran arsitektur sistem secara singkat namun menyeluruh dan menggambarkan keseluruhan fase yang ada, tahapan serta proses yang dilakukan dapat dilihat pada gambar 2.



Gambar 2. Arsitektur Sistem

Arsitektur sistem yang dibangun di penelitian ini dibagi menjadi 3 fase yaitu fase ekstraksi data, pembentukan dataset dan fase klasifikasi data dengan hasil belajar dari dataset yang telah dibentuk pada fase sebelumnya.

A. Fase Ekstraksi Data

Pada fase ekstraksi data, sistem akan mencari informasi berita sepak bola tersebut melalui crawling situs goal.com/en dan fifa.com. Dari setiap halaman yang berhasil dikunjungi oleh crawler, akan didapatkan halaman-halaman berita. Setiap halaman berita yang didapatkan pada tahap ini masih berupa sebuah dokumen HTML yang memuat banyak tag-tag

HTML yang perlu dibersihkan untuk mendapatkan informasi penting dari berita tersebut seperti judul, isi dan informasi-informasi lain terkait berita.

```
<article>
<a href="/en/news/12/spain/2013/10/02/4305154/real-madrid-handed-bale-blow">

</a>
<h3><a href="/en/news/12/spain">Spain</a></h3>
<h2><a href="/en/news/12/spain/2013/10/02/4305154/real-madrid-handed-bale-blow">Bale out of Copenhagen clash</a></h2>
<!-- module:SocialShareIcons -->
<div class="module module-social-share-icons dark" data-role="share-icons" data- url="http://www.goal.com/en/news/12/spain/2013/10/02/4305154/real-madrid-handed-bale-blow" data-title="Bale out of Copenhagen clash" data-description="The 24-year-old has suffered a stop-start beginning to his Blancos career and will miss the Liga side's latest European tie" data-image="http://u.goal.com/322600/322604_hp.jpg">
```

Gambar 3. Contoh Hasil Proses Fetching

Selanjutnya pada dokumen HTML yang diperoleh dilakukan ekstraksi informasi id, judul, sub judul, tanggal, tag, isi berita, sumber, hingga link gambar terkaitnya dengan menggunakan regex, hingga diperoleh informasi seperti yang ditunjukkan pada gambar 4 berikut ini.

```
Kode Berita : FF2261338
Judul Berita : Matri moved to Fiorentina to fix injury crisis
Sub Judul Berita : Alessandro Matri has been loaned out to Fiorentina from rivals AC Milan on a six-month contract to help fill La Viola's void up top with strikers Gius...
Isi Berita : Italian Serie A side Fiorentina moved to resolve an injury crisis affecting their strikers by signing Italian international striker Alessandro Matri on loan from rivals left by injuries to Giuseppe Rossi, leading scorer in Serie A with 14 goals, and German international Mario Gomez."There are some minor details to eke out and then I will be available for the coach,"
Pembuat Berita : Fifa
Sumber : Fifa
Tag Berita : Italy
URL foto Berita : www.fifa.com/mm/photo/tournament/competition/01/69/73/08/1697308_small-1nd.jpg
Tanggal Berita : 2013-10-10
```

Gambar 4. Contoh Hasil Ekstraksi Informasi

Proses selanjutnya yang dilakukan adalah melakukan ekstraksi setiap teks komentar yang diberikat terhadap artikel berita terkait. Proses yang dilakukan sama seperti proses ekstraksi informasi berita, yaitu dengan menggunakan regex. Setelah berhasil mendapatkan informasi tersebut, sistem akan melakukan filter untuk mendapatkan komentar yang menggunakan bahasa Inggris saja. Untuk membedakan komentar tersebut menggunakan bahasa Inggris atau tidak, cara yang akan digunakan yaitu dengan menggunakan language detector API.

B. Fase Pembentukan Dataset

Karena Naïve Bayes merupakan algoritma yang termasuk dalam kategori supervised learning, maka terlebih dahulu dilakukan pembentukan dataset. Dataset ini penting digunakan untuk proses training algoritma Naïve Bayes itu sendiri. Penggunaan algoritma Naïve Bayes juga mengharuskan statistika persebaran dataset untuk setiap kata berada pada posisi yang seimbang.

Dataset yang digunakan pada penelitian ini diambil dari hasil ekstraksi komentar yang didapat dari situs yang sudah ditentukan yaitu fifa.com dan goal.com/en. Kemudian dataset tersebut diberi label secara manual yaitu positif, negatif atau netral. Positif dilambangkan dengan P, Negatif dilambangkan

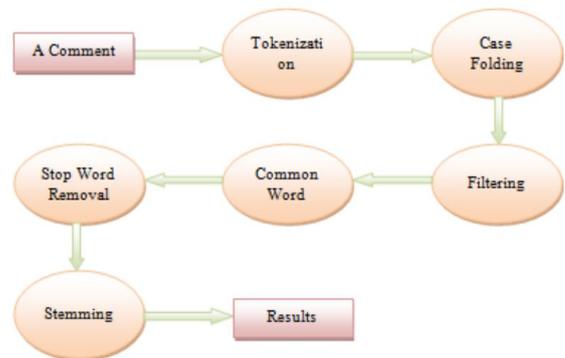
dengan N, sedangkan Netral dilambangkan dengan O. Dataset yang digunakan berjumlah 20.000 komentar. Selanjutnya dataset ini akan diproses melalui preprocessing. Pada program sesungguhnya dataset yang digunakan berjumlah 500 komentar. Contoh bentuk dataset yang digunakan sebagai data training dapat dilihat pada gambar 5 berikut ini.

```
O Well said.
P Glad to see an Englishman speaking the truth.
N wanker would be appropriate
P As much as I hate this guy, I gotta agree with him for once!
P lmao i like this guy
N Barton is a twat but he's spot on here about the national team lol. However his criticism of SAF is uncalled for.
P lol love this guy says what most people think
N England is a retarded footballing nation at the core. seriously..... no English manager has ever won the PL.... no other league is that depended on foreigners
O When was england ever good ?
O When they could cheat
N Ferguson is not from England...
O Barton is so raw, he is just straight up badass
O true, how many British Players are being Exported to other countries? definitely not as much as the Spanish or Brazilian or even mexican players.
N And you would know crap wouldn't you Barton.
N Barton is an ass, that being said he kinda has point even if he just is an ass
P wow. Joey is such a fearless lad. finally he made great point and I totally agree with him. England need a change seriously.
P Good bang
```

Gambar 5. Contoh Dataset

C. Fase Klasifikasi Data

Proses Preprocessing dimulai dengan melakukan proses tokenisasi pada sebuah komentar. Proses tokenisasi adalah proses untuk memecah suatu komentar menjadi sekumpulan kata. Biasanya proses ini akau memecah suatu dokumen berdasarkan karakter spasi, atau tanda koma.



Gambar 6. Preprocessing

Proses selanjutnya case folding yaitu proses yang digunakan untuk penyamaan case huruf. Filtering adalah proses untuk menghilangkan karakter yang tidak digunakan dalam penentuan sebuah kelas sentimen. Karakter-karakter yang akan dihilangkan seperti kata yang mengandung perulangan huruf lebih dari tiga, tanda baca, symbol.

Common word dan stop word removal adalah proses untuk menghilangkan kata-kata yang tidak dipakai dalam proses penentuan sebuah kelas sentimen. Untuk common word removal, kata-kata yang dihilangkan adalah kata-kata yang ada hubungan dengan topik penelitian, seperti nama pemain bola, nama stadion, nama pelatih, nama klub, nama negara, singkatan dan lain-lain. Sedangkan untuk proses stop word removal, kata-kata yang dihilangkan antara lain kata hubung, kata ganti orang, kata benda, dan lain-lain.

Stemming adalah proses yang digunakan untuk mengembalikan sebuah kata menjadi kata dasarnya. Metode stemming yang akan digunakan yaitu porter stemmer [10]. Metode ini ditemukan oleh Martin F. Porter. Metode ini paling sering digunakan dalam proses stemming untuk bahasa Inggris.

Setelah proses preprocessing selesai, proses training data baru dilakukan. Berikut adalah contoh hasil proses training data.

TABEL 2
DAFTAR STATISTIK DAN PROBABILITAS

Words	Jumlah			Probabilitas		
	Pos	Neg	Net	Pos	Neg	Net
well	3	3	0	0.00300	0.00207	0.00088
said	3	4	1	0.00300	0.00259	0.00176
glad	1	0	0	0.00150	0.00051	0.00088
hate	1	3	0	0.00150	0.00207	0.00088
gotta	1	0	0	0.00150	0.00051	0.00088
agree	3	0	0	0.00300	0.00051	0.00088
like	4	7	3	0.00375	0.00414	0.00353
love	3	2	0	0.00300	0.00155	0.00088
retart	0	1	0	0.00075	0.00103	0.00088

Setelah melakukan perhitungan proses training dan perhitungan probabilitas dari training data tersebut dapat dilakukan proses pengklasifikasian sentimen. Berikut adalah contoh proses pengklasifikasian sentimen.

- Kata yang ingin diklasifikasi:
i don't like barton and england. Both of them are retarded.
- Kata setelah preprocessing: like dan retart
- Perhitungan Probabilitas Positif:
 $0.00375 * 0.00075 = 0.0000028125$
- Perhitungan Probabilitas Negatif:
 $0.00414 * 0.00103 = 0.0000042642$
- Perhitungan Probabilitas Netral:
 $0.00353 * 0.00088 = 0.0000031064$

Dari perhitungan ketiga probabilitas tersebut, kata yang ingin diklasifikasi dapat digolongkan ke sentimen negatif karena nilai probabilitas negatif lebih besar dari nilai probabilitas positif dan probabilitas netral.

IV. INPUT OUTPUT SISTEM

Pada tahap ini akan dijelaskan mengenai input dan output dari sistem ini. Sistem akan diberi inputan yang kemudian akan diproses untuk mendapatkan output. Berikut akan dibahas input dan output sistem beserta contoh-contoh yang relevan.

A. Input

Input yang diperlukan sistem adalah kalimat dalam bahasa Inggris yang baik dan benar tentang komentar berita sepak bola. Kalimat yang diberikan tidak boleh memiliki singkatan, perulangan huruf yang tidak sesuai atau kata yang tidak baku. Contoh kalimat input yang baik:

- yes you are the best but only for barcelona because you not going anywhere but christiano ronaldo proof he is the best alltime anywhere any place
- I love messi ! No one can beat him!

B. Output

Sistem memberikan beberapa jenis output sehubungan

dengan input teks yang diberikan oleh pengguna. Pertama, output berupa penggolongan bahasa yang didapat menggunakan Language Detection API. Ada kemungkinan sistem mendapatkan satu atau lebih jenis penggolongan bahasa. Output yang kedua adalah pengklasifikasian input teks berdasarkan kedua algoritma yang digunakan. Untuk algoritma naive bayes, input teks akan digolongkan ke sentimen positif, negatif, atau netral. Untuk algoritma Baseline, input teks hanya digolongkan ke sentimen positif atau negatif

V. UJICOBA

Uji coba sistem meliputi uji coba akurasi output yang dilakukan dengan menghitung hasil akurasi, precision dan recall [11] hasil identifikasi 150 komentar berita yang didapatkan baik melalui crawling atau penambahan komentar oleh member. Dari perhitungan macroaveraged untuk setiap kelas, didapatkan akurasi 93,06%, presisi 81,90%, dan recall 92,67% untuk kelas positif. Dari perhitungan kelas negatif didapatkan akurasi 87,73%, presisi 96,29%, dan recall 83,63%. Dari perhitungan kelas netral didapatkan akurasi 92,26%, presisi 64,44%, dan recall 90,37%.

VI. KESIMPULAN

Kesimpulan yang diperoleh saat penelitian ini dari awal hingga akhir adalah.

1. Proses crawling yang digunakan untuk mendapatkan berita dan komentar berita sangat membantu dalam penambahan konten website, tetapi banyak sekali komentar berita yang diperoleh kurang cocok untuk proses klasifikasi. Oleh karena itu dibutuhkan penyaringan komentar sehingga didapatkan komentar yang lebih baik.
2. Agar hasil proses preprocessing menjadi lebih baik, sistem membutuhkan daftar stopword dan commonword yang lebih banyak. Semakin baik hasil dari proses preprocessing akan membantu proses klasifikasi data.
3. Untuk meningkatkan kinerja dari proses klasifikasi sentiment diperlukan training data yang lebih banyak. Dengan hasil training yang lebih banyak diharapkan akan mendapatkan akurasi yang lebih baik.
4. Penggunaan Language Detection API untuk pendeteksian bahasa sangat membantu dalam penentuan komentar yang digunakan untuk proses klasifikasi, tetapi sebagian pendeteksian yang dilakukan tidak menghasilkan bahasa yang sesuai.

DAFTAR PUSTAKA

- [1] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168–177, doi: 10.1145/1014052.1014073.
- [2] B. Liu and others, "Sentiment analysis and subjectivity.," *Handb. Nat. Lang. Process.*, vol. 2, no. 2010, pp. 627–666, 2010.
- [3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," *arXiv Prepr. cs/0205070*, 2002.
- [4] B. Pang, L. Lee, and others, "Opinion mining and sentiment analysis.," *Found. Trends@in Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [5] M. A. Rahman, H. Budiarto, and E. I. Setiawan, "Aspect Based Sentimen Analysis Opini Publik Pada Instagram dengan Convolutional Neural Network," *J. Intell. Syst. Comput.*, vol. 1,

- no. 2, pp. 50–57, 2019.
- [6] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, “Predicting the political alignment of twitter users,” in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, 2011, pp. 192–199.
- [7] A. Makazhanov, D. Rafiei, and M. Waqar, “Predicting political preference of Twitter users,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–15, 2014.
- [8] N. W. S. Saraswati, “Text mining dengan metode naive bayes classifier dan support vector machines untuk sentiment analysis,” *Univ. UDAYANA, Tek. Elektro. Denpasar Univ. UDAYANA*, 2011.
- [9] S. I. Wang and C. D. Manning, “Baselines and bigrams: Simple, good sentiment and topic classification,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 90–94.
- [10] M. F. Porter, “An algorithm for suffix stripping,” *Program*, 1980.
- [11] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*, vol. 39. Cambridge University Press Cambridge, 2008.