

# Analisis Sentimen Pasar Saham dari Berita Keuangan menggunakan Algoritma Pencocokan String

Raihan Astrada Fathurrahman (13519113)

Program Studi Teknik Informatika  
Sekolah Teknik Elektro dan Informatika  
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung  
E-mail (gmail): raihan.astrada@gmail.com

**Abstract**—Dewasa ini, terdapat berbagai macam instrumen keuangan yang dapat digunakan untuk menambah nilai aset keuangan, salah satunya adalah saham. Saham banyak diminati oleh masyarakat karena saham menawarkan tingkat keuntungan yang signifikan dengan modal yang relatif terjangkau. Salah satu cara mendapatkan keuntungan dari saham adalah dengan memperoleh keuntungan dari kenaikan harga saham. Terdapat berbagai faktor yang dapat mempengaruhi perubahan harga saham, salah satunya adalah sentimen pasar saham. Sentimen pasar saham ini dapat kita peroleh dari berita keuangan yang beredar. Pada makalah ini, akan dibahas mengenai penggunaan algoritma pencocokan string untuk scraping berita keuangan dan pengolahan data hasil scraping menjadi sebuah sentimen saham. Hasil analisis sentimen ini dapat digunakan dalam menentukan opini investor atas saham atau aset tertentu.

**Keywords**—analisis sentimen; berita keuangan; saham; pencocokan string;

## I. PENDAHULUAN

Dewasa ini, terdapat berbagai macam instrumen keuangan yang dapat kita gunakan untuk menambah nilai aset keuangan yang kita miliki. Beberapa instrumen keuangan yang kita sering ketahui yaitu Saham, Reksadana, Surat Utang (Obligasi), dan Deposito. Dari berbagai macam instrumen keuangan yang ada, salah satu instrumen keuangan yang saat ini diminati oleh banyak masyarakat adalah Saham. Alasan saham banyak diminati oleh masyarakat karena saham menawarkan tingkat keuntungan yang signifikan dengan modal yang relatif terjangkau oleh masyarakat.

Saham adalah bukti kepemilikan seseorang terhadap kekayaan suatu perusahaan. Saat seseorang membeli saham suatu perusahaan, maka orang tersebut sudah mempunyai hak atas suatu aset dan juga pendapatan yang diperoleh dari perusahaan tersebut sesuai dengan porsi saham yang dibelinya serta berhak hadir dalam rapat umum pemegang saham (RUPS) [3].

Saham dijualbelikan pada suatu pasar yang dinamakan Bursa Efek Indonesia (BEI). Pada BEI saham yang diperjualbelikan juga memiliki perubahan harga. Harga saham merupakan harga yang terjadi di bursa pada saat tertentu berdasarkan permintaan dan penawaran yang terjadi. Harga inilah yang kemudian menjadi salah satu cara untuk mendapatkan keuntungan yaitu dengan *Capital Gain*. Capital

Gain merupakan keuntungan yang diperoleh dari kenaikan nilai aset, dalam hal ini seseorang membeli saham pada harga rendah kemudian terjadi peningkatan harga saham dan keuntungan dari selisih tersebut merupakan *Capital Gain*.



Gambar 1 Saham di Bursa Efek Indonesia

Sumber:

<https://ekonomi.kompas.com/read/2017/11/03/221703226/kenaikan-indeks-harga-saham-di-bursa-efek-indonesia-tertinggi-di-dunia>

Salah satu faktor yang mempengaruhi perubahan harga saham yaitu adanya sentimen pasar saham. Sentimen ini berkaitan dengan psikologi trading yang kemudian dapat memengaruhi perilaku pemegang saham dalam melakukan jualbeli. Ketika sentimen pasar saham positif maka pemegang saham akan cenderung untuk melakukan pembelian, begitu pula sebaliknya. Volatilitas inilah yang kemudian dapat menggerakkan harga saham.

Untuk mengetahui sentimen pasar saham pada BEI, salah satu caranya adalah dengan melakukan analisis sentiment pasar saham yang diperoleh melalui berita-berita keuangan. Untuk membatasi banyaknya data, berita yang diambil merupakan berita yang diterbitkan antara tanggal 4 Mei 2021 sampai 11 Mei 2021 dan data saham yang diambil merupakan saham-saham yang terdapat pada index LQ45 periode Februari-Juli 2021. Hasil analisis sentimen ini, digunakan dalam menentukan opini investor atas saham atau aset tertentu

## II. TEORI DASAR

### A. Analisis Sentimen

Analisis sentimen atau sering disebut juga *opinion mining* adalah deteksi sikap-sikap terhadap objek atau orang [Beineke et al. 2014]. Analisis sentimen dapat digunakan untuk mendapatkan persentase sentimen terhadap seseorang, perusahaan, institusi, produk atau pada sebuah kondisi tertentu. Nilai dari analisis sentimen bisa dipecah menjadi 3 yakni, sentimen positif, sentimen negatif dan sentimen netral atau diperdalam lagi sehingga dapat menemukan siapa atau kelompok yang menjadi sumber sentimen positif atau sentimen negatif [4].

Analisis sentimen dilakukan untuk mengetahui pendapat terhadap sebuah masalah, atau dapat juga digunakan untuk mengidentifikasi kecenderungan suatu hal di pasar. Terdapat berbagai macam manfaat dari analisis sentiment, beberapa diantaranya yaitu dapat digunakan untuk memperoleh gambaran secara umum dari persepsi masyarakat terhadap kualitas suatu layanan, pemantauan sebuah produk dari suatu industri, prediksi suatu penjualan, hingga alat bantu pengambilan keputusan yang dapat digunakan oleh tokoh politik maupun investor.

### B. VADER

VADER (*Valence Aware Dictionary and sEntiment Reasoner*) merupakan sebuah kaskas leksikon dan analisis sentiment yang spesifik dibuat untuk mengekstraksi sentimen yang diekspresikan di media sosial. Kaskas ini terdapat pada bahasa pemrograman *Python, Java, JavaScript, PHP, Scala, C#, Rust, Go, dan R* serta mampu mendukung analisis sentimen untuk kalimat berbahasa Inggris beserta emoticon (tulisan tipografi yang merepresentasikan ekspresi wajah).

Penilaian analisis sentimen pada VADER dapat dikategorikan menjadi 4 bagian yaitu *pos, neg, neu, dan compound*. *Pos, neu, dan neg* merupakan nilai rasio proporsi dari teks yang termasuk dalam setiap kategorinya, yang mana *pos* berarti teks bersifat positif, *neu* berarti teks bersifat netral, dan *neg* berarti teks bersifat negatif. Hal ini dapat digunakan untuk menganalisis konteks dari suatu kalimat. Sedangkan nilai *compound* dihitung dengan menjumlahkan skor valensi setiap kata dalam leksikon, disesuaikan dengan aturan VADER (mempertimbangkan pengurutan kata, penguat bentuk kata, penguat tanda baca, sensitivitas konjungsi yang kontras), dan kemudian dinormalisasi menjadi antara -1 (negatif paling ekstrim) dan +1 (positif paling ekstrim). Nilai *compound* ini dapat digunakan sebagai satuan standar untuk mengklasifikasikan kalimat dengan ketentuan *positive* untuk  $compound \geq 0.05$ , *negative* untuk  $compound \leq -0.05$  dan *neutral* untuk  $-0.05 < compound < 0.05$  [2].

```
VADER is very smart, handsome, and funny.-----
{'pos': 0.701, 'compound': 0.8545, 'neu': 0.299, 'neg': 0.0}
VADER is VERY SMART, handsome, and FUNNY.-----
{'pos': 0.754, 'compound': 0.9227, 'neu': 0.246, 'neg': 0.0}
VADER is VERY SMART, handsome, and FUNNY!!!-----
{'pos': 0.767, 'compound': 0.9342, 'neu': 0.233, 'neg': 0.0}
```

```
Today SUX!-----
{'pos': 0.0, 'compound': -0.5461, 'neu': 0.221, 'neg': 0.779}
Today only kinda sux! But I'll get by, lol-----
{'pos': 0.317, 'compound': 0.5249, 'neu': 0.556, 'neg': 0.127}
Make sure you :) or :D today!-----
{'pos': 0.706, 'compound': 0.8633, 'neu': 0.294, 'neg': 0.0}
Catch utf-8 emoji such as 🍷 and 🍷 and 🤪-----
{'pos': 0.279, 'compound': 0.7003, 'neu': 0.721, 'neg': 0.0}
```

### C. Pencocokan String

String merupakan tipe data yang merupakan urutan kumpulan alphanumeric karakter maupun symbol yang berdekatan. Pada string terdapat beberapa definisi seperti awalan (*prefix*) dan akhiran (*suffix*). Misalkan S merupakan sebuah string dengan panjang ukuran k sebagai berikut,

$$S = X_1, X_2, X_3, \dots, X_k$$

maka *prefix* dari S adalah *substring* dari  $S[0..n]$  dan *suffix* dari S adalah *substring* dari  $S[n..k]$  dengan n merupakan indeks antara 0 sampai dengan k. Berikut merupakan contoh string beserta *prefix* dan *suffix* nya.

- $S = \text{contoh}$
- Semua kemungkinan *Prefix* dari S:  
"c", "co", "con", "cont", "conto", "contoh"
- Semua kemungkinan *Suffix* dari S:  
"h", "oh", "toh", "ntoh", "ontoh", "contoh"

Pencocokan string merupakan persoalan untuk mencari lokasi pertama kata yang bersesuaian dengan *pattern* di dalam suatu *text*, yang mana *text* dan *pattern* merupakan suatu string dengan *text* merupakan string yang panjangnya jauh lebih besar dibanding *pattern* yang akan dicari di dalam teks.

### D. Algoritma Pencocokan String

Algoritma pencocokan string merupakan algoritma yang digunakan untuk menyelesaikan persoalan pencocokan string. Saat ini, terdapat berbagai macam algoritma pencocokan string, seperti Algoritma *Brute Force*, Algoritma *Knuth-Morris-Pratt* (KMP), Algoritma *Boyer-Moore* (BM), dan juga dengan *Regular Expression* (Regex). Algoritma pencocokan string yang digunakan pada makalah ini adalah dengan Regex karena data yang ingin discraping dari web sulit untuk dilakukan secara *exact matching* dan lebih mudah untuk melakukan *regex matching* untuk mengambil data-data yang diperlukan serta regex dapat digunakan untuk memodifikasi data yang didapatkan.

Regex merupakan sebuah string yang merepresentasikan sebuah pola pencarian sehingga dapat membantu kita untuk melakukan pencocokan string serta memanipulasi teks. Berikut ini beberapa notasi yang digunakan dalam regex.

Notasi	Penjelasan
.	Karakter apapun kecuali newline

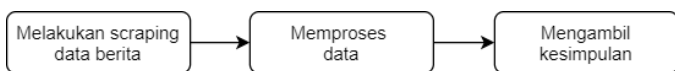
<code>\w \d \s</code>	Kata, Digit, Spasi
<code>\W \D \S</code>	Bukan kata, bukan digit, bukan spasi
<code>[abc]</code>	Bisa a, b, atau c
<code>[^abc]</code>	Bukan a, b, atau c
<code>[a-g]</code>	Karakter antara a & g
<code>^abc\$</code>	Awal / Akhir dari string
<code>(expr)</code>	Mengelompokkan token pada <i>expr</i>
<code>(?:expr)</code>	Mengelompokkan token pada <i>expr</i> tanpa membuat capture group
<code>(?=expr)</code>	Diikuti oleh <i>expr</i>
<code>(?!expr)</code>	Tidak diikuti oleh <i>expr</i>
<code>*</code>	0 atau lebih token
<code>+</code>	1 atau lebih token
<code>?</code>	0 atau 1 token
<code>{n}</code>	Tepat n token berurut
<code>{n,}</code>	N atau lebih token berurut
<code>{m,n}</code>	Antara m dan n token berurut
<code>??, *?, +?, {n}?, etc.</code>	Untuk mencari token berdasarkan jumlah yang sesedikit mungkin
<code>ab cd</code>	Bisa ab atau cd

Tabel 1 Notasi Regex

Sumber: Penulis

### III. PEMBAHASAN

Secara garis besar, terdapat 3 buah tahapan dalam melakukan analisis sentimen pasar saham dari berita keuangan. Tahap yang pertama adalah scraping data dari *website* berita keuangan yang ingin diambil data beritanya. Tahap kedua adalah memproses data yang telah discrap. Tahap yang terakhir adalah menarik kesimpulan dari data yang telah didapatkan. Berikut adalah gambar alur proses yang dilakukan dalam menganalisis sentimen pasar saham.



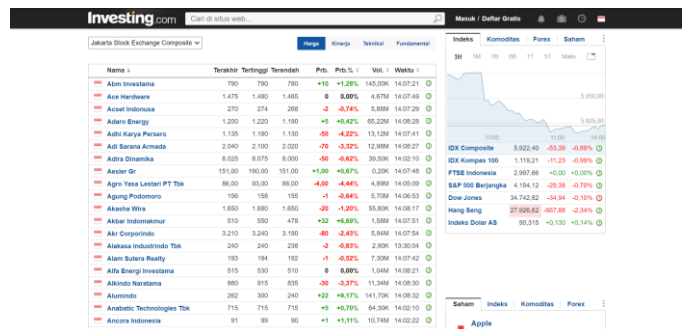
Gambar 2 Alur Proses Analisis Sentimen Pasar Saham

Sumber: Penulis

#### A. Scraping Data Berita

Scraping adalah suatu metode mengekstraksi data pada sebuah *website*. Proses scraping ini dapat dilakukan secara manual maupun otomatis dengan menggunakan sebuah tool. Pada makalah ini, scraping yang dilakukan yaitu untuk mengekstraksi berita-berita keuangan yang terdapat pada suatu *website* dengan menggunakan bantuan *library* BeautifulSoup

pada Bahasa pemrograman Python. Halaman *website* yang digunakan sebagai target scraping adalah halaman [id.investing.com](https://id.investing.com/equities/indonesia) (<https://id.investing.com/equities/indonesia>). Berita yang diambil merupakan berita-berita yang diterbitkan antara tanggal 4 Mei 2021 sampai 11 Mei 2021 pada halaman pertama saham-saham yang terdapat pada index LQ45 periode Februari-Juli 2021.



Gambar 3 Tampilan halaman *website* id.investing.com

Sumber: Screenshot Penulis dari *web* id.investing.com

Pada proses scraping, terdapat beberapa regex yang digunakan untuk mengambil dan memodifikasi data-data yang diperlukan. Regex-regex tersebut adalah sebagai berikut

Regex	Penjelasan
<code>&gt;[\w\s(&amp;#amp;#x27;)]*&lt;/a&gt;</code>	Digunakan untuk mengambil nama saham. Nama saham terletak diantara tag <code>&lt;a&gt;</code> pada <i>website</i>
<code>href="[\w-\/]*\"</code>	Digunakan untuk mengambil unique url saham. Unique url disimpan diantara href pada <i>website</i>
<code>(\d{2})[/-](\d{2})[/-](\d{4})</code>	Digunakan untuk mengambil tanggal dengan format dd/mm/yyyy
<code>title=\"[\w\s!@#%&amp;^&amp;()-\`.,+/*]*\"</code>	Digunakan untuk mengambil headline dari berita. Headline berita disimpan diantara title pada <i>website</i>
<code>&lt;/a&gt;&gt;</code>	Digunakan untuk menghapuskan tag <code>&lt;a&gt;</code> pada nama saham
<code>&amp;#amp;</code>	Digunakan untuk mensubstitusi <code>&amp;#amp;</code> menjadi <code>&amp;</code> . Hal ini karena symbol <code>'&amp;'</code> pada HTML tertulis sebagai <code>'&amp;#amp;#x27;</code>
<code>href=\\\"</code>	Digunakan untuk menghapuskan href pada unique url saham
<code>title=\\\"</code>	Digunakan untuk

	menghapuskan title pada headline berita
--	---

**Tabel 2** Regex yang digunakan pada proses scraping data

Sumber: Penulis

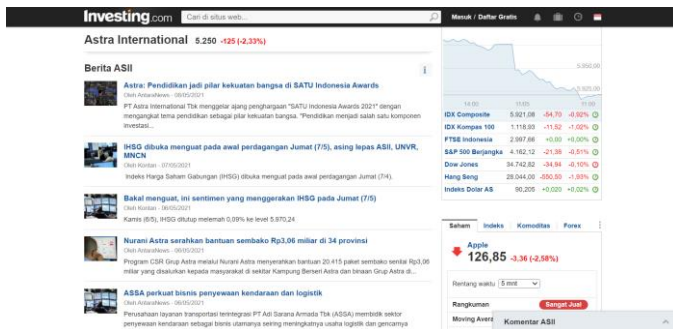
Berikut juga beberapa contoh hasil pemakaian modifikasi string menggunakan regex.

Awal	Akhir	Penjelasan
>Pudjadi & Sons</a>	Pudjadi & Sons	Menghapuskan tag <a> kemudian mengubah '&' menjadi symbol '&'
href="/equities/ahap-tbk"	/equities/ahap-tbk	Menghapuskan href dan tanda kutip ("")
title="IHSG naik 0,17% sepanjang April, ini saham-saham dengan net buy & sell terbesar asing"	IHSG naik 0,17% sepanjang April, ini saham-saham dengan net buy & sell terbesar asing	Menghapuskan title dan tanda kutip ("") serta mengubah '&' menjadi symbol '&'

**Tabel 3** Contoh penggunaan regex untuk memodifikasi

Sumber: Penulis

Proses scraping data dimulai dengan mengakses tampilan utama *website* (<https://id.investing.com/equities/indonesia>). Dapat dilihat dari Gambar 3 bahwasanya tampilan utama hanya memuat nama-nama saham. Hal tersebut karena berita-berita pada *website* terletak pada unique url dari setiap sahamnya. Pada website id.investing.com unique url ini disimpan sebagai alamat yang dituju ketika meng-klik nama-nama saham. Maka untuk mendapatkan berita dari setiap sahamnya, hal yang dilakukan pertama kali adalah mengakses konten tampilan utama *website*. kemudian mengambil serta memodifikasi data-data nama dan unique url website dengan menggunakan regex-regex yang telah dijelaskan pada tabel 2 dan tabel 3.



**Gambar 4** Tampilan Berita Keuangan Saham Astra Internasional pada *website* id.investing.com

Sumber: Screenshot Penulis dari web id.investing.com

Setelah mendapatkan unique url untuk setiap sahamnya, maka hal yang dilakukan adalah mengakses tampilan berita keuangan kemudian mengekstrak berita-berita dari setiap unique url untuk setiap sahamnya. Proses pengambilan data

headline dan tanggal berita ini juga menggunakan regex yang telah dijelaskan pada tabel 2. Tanggal digunakan sehingga berita yang didapatkan hanya berita yang memiliki jangka waktu satu minggu kebelakang dari waktu menjalankan program. Dari proses ekstraksi berita, didapatkan 239 berita dari 45 saham. Berikut beberapa data berita hasil ekstraksinya.

[
['Ace Hardware', 'Penjualan Lesu, Laba Ace Hardware Turun 27,9% di 2020'],
['Ace Hardware', 'Ekspansif, Ace Hardware Buka Toko Keempat di Tahun 2021'],
['Ace Hardware', 'Ace Hardware Indonesia (ACES) akan buka gerai baru di Depok, Jawa Barat'],
['Ace Hardware', 'IHSG menguat 0,20% ke 5.975 pada sesi I hari ini, asing beli BBRI, TBIG, BBKA'],
['Adaro Energy', 'Prediksi IHSG hari Jumat 7/5 bisa menembus 6.000, ini 3 saham yang bakal naik harga'],
...
['Bank Tabungan Pensiunan', 'Kinclongnya Kinerja 2 Bank Syariah Ungguli Bank Konvensional'],
['Bank Tabungan Pensiunan', 'IHSG dibuka menguat pada awal perdagangan Jumat (7/5), asing lepas ASII, UNVR, MNCN']
]

## B. Processing Data Berita

Setelah mendapatkan berita dari setiap saham, tahap yang dilakukan selanjutnya adalah memproses/mengolah data berita yang didapatkan. Pemrosesan ini terdiri atas menerjemahkan berita, mencari analisis sentimen, dan memvisualisasikan data.

Proses pertama yaitu menerjemahkan berita. Proses ini perlu dilakukan karena pada *library* yang digunakan untuk melakukan analisis sentimen, VADER, belum mendukung analisis sentimen untuk Bahasa Indonesia sehingga perlu untuk menerjemahkan berita Bahasa Indonesia ke Bahasa Inggris. Untuk membantu menerjemahkan berita, penulis menggunakan *library* TextBlob. TextBlob merupakan salah satu *library* pada python yang digunakan untuk memproses data tekstual. TextBlob juga menyediakan fungsi untuk translate yang mengakses API *Google Translate* dalam penggunaannya.

Setelah seluruh berita berhasil diterjemahkan proses selanjutnya adalah melakukan analisis sentimen. Analisis sentimen yang digunakan menggunakan bantuan *library* VADER. Hasil analisis sentiment menggunakan VADER akan menghasilkan 4 buah hasil yaitu *pos*, *neg*, *neu*, dan *compound*. Pada makalah ini hasil sentimen yang akan digunakan adalah *compound* yang merupakan agregasi untuk tiga buah hasil lainnya (*pos*, *neg*, dan *neu*). Sebelum hasil analisis sentiment divisualisasikan, akan dilakukan pengelompokan data terlebih dahulu berdasarkan nama saham dan kemudian dilakukan



materiil, serta kepada teman-teman yang senantiasa mendukung dan memberikan motivasi dalam membuat makalah ini.

#### REFERENCES

- [1] Bohmian. 2020. Sentiment Analysis of Stocks from Financial News using Python. <https://towardsdatascience.com/sentiment-analysis-of-stocks-from-financial-news-using-python-82ebdcefb638>. Diakses 8 Mei 2021 pada pukul 12.16 WIB.
- [2] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
- [3] Ibnuismail. 2020. Pengertian Saham: Definisi, Jenis, Manfaat, dan Risiko yang Harus Ditanggung. <https://accurate.id/ekonomi-keuangan/pengertian-saham/#:~:text=Sebenarnya%20pengertian%20saham%20adalah%20suatu.bagian%20kepemilikan%20dari%20suatu%20perusahaan.&text=Pengertian%20saham%20yang%20lainnya%20bisa.bagian%20bentuk%20kepemilikan%20suatu%20perusahaan>. Diakses 10 Mei 2021 pada pukul 22.49
- [4] Suyanto, Data Mining untuk klasifikasi dan klusterisasi data, Penerbit INFORMATIKA. Bandung. 2017

- [5] <https://informatika.stei.itb.ac.id/~rinaldi.munir/Stmik/2020-2021/Pencocokan-string-2021.pdf>. Diakses 11 Mei 2021 pada pukul 9.15
- [6] <https://informatika.stei.itb.ac.id/~rinaldi.munir/Stmik/2018-2019/String-Matching-dengan-Regex-2019.pdf>. Diakses 11 Mei 2021 pada pukul 9.30

#### PERNYATAAN

Dengan ini saya menyatakan bahwa makalah yang saya tulis ini adalah tulisan saya sendiri, bukan saduran, atau terjemahan dari makalah orang lain, dan bukan plagiasi.

Jakarta, 11 Mei 2021



Raihan Astrada Fathurrahman  
13519113