**LAPORAN *ASPECT EXTRACTION* DENGAN MENGGUNAKAN METODE *PATTERN BASED***

mochamad rakha luthfi f – 1301164573, Joshua Panjaitan- 1301164044

**Abstrak**

Perkembangan teknologi saat ini begitu cepat, terutama dalam pertumbuhan jumlah data. Peningkatan penggunaan akun social media dan transaksi online masih menjadi komoditas panas dalam perkembangan teknologi saat ini. Dalam bidang traksaksi online, ulasan produk merupakan data yang tidak bisa ditinggalkan begitu saja, perlu dilakukan pengolahan untuk dapat menganalisis pasar didalam perkembangan teknologi saat ini. Sentimen analisis merupakan cabang ilmu yang mengolah data teks dan memberikan analisis terhadap data tersebut. Dalam sentimen analis terdapat tahapan ekstraksi aspek yang dimana tujuannya untuk menentukan aspek dan opini yang melekat pada suatu kalimat, yang nantinya akan digunakan dalam proses selanjutnya. Terdapat banyak metode untuk menemukan pasangan aspek dan opini, salah satunya adalah metode *rule based*, metode ini terbilang sederhana dan cenderung memberikan hasil yang tidak mengecewakan. Pada penelitian ini akan dilakukan *aspect extraction* dengan metode *rule based* untuk menentukan pasangan aspek dan opini, kemudian dilakukan pengujian untuk menentukan hasil akurasi dari metode tersebut.

## **Pengantar**

*Aspect Extraction* merupakan salah satu tahapan dalam Sentimen Analisis yang dimana tujuannya adalah untuk mendapatkan pasangan aspek dan opini yang terdapat dalam sebuah kalimat. Pasangan aspek dan opini yang telah di ekstrak nantinya akan sangat berguna untuk tahapan selanjutnya dalam analisis sentimen, sehingga perlu untuk menemukan pasangan aspek dan opini yang sesuai untuk proses selanjutnya. Dalam menentukan aspek dalam sebuah kalimat terdapat beberapa cara yang bisa digunakan, salah satunya adalah metode *Pattern Rule Based.* Metode berbasis aturan yang telah didefinisikan sebelumnya oleh peneliti sebelumnya untuk menemukan pasangan aspek opini. Metode ini relatif mudah untuk diimplementasikan namun memiliki beberapa kekurangan dalam menemukan aspek yang sesuai karena hanya terbatas dalam aturan yang relatif kaku. Dalam penelitian ini metode *pattern based* dipilih untuk diimplementasikan dikarenakan kemudahan dalam implementasi dalam kasus yang tidak terlalu kompleks, dengan dataset ulasan camera Canon G3.

## **Metode**

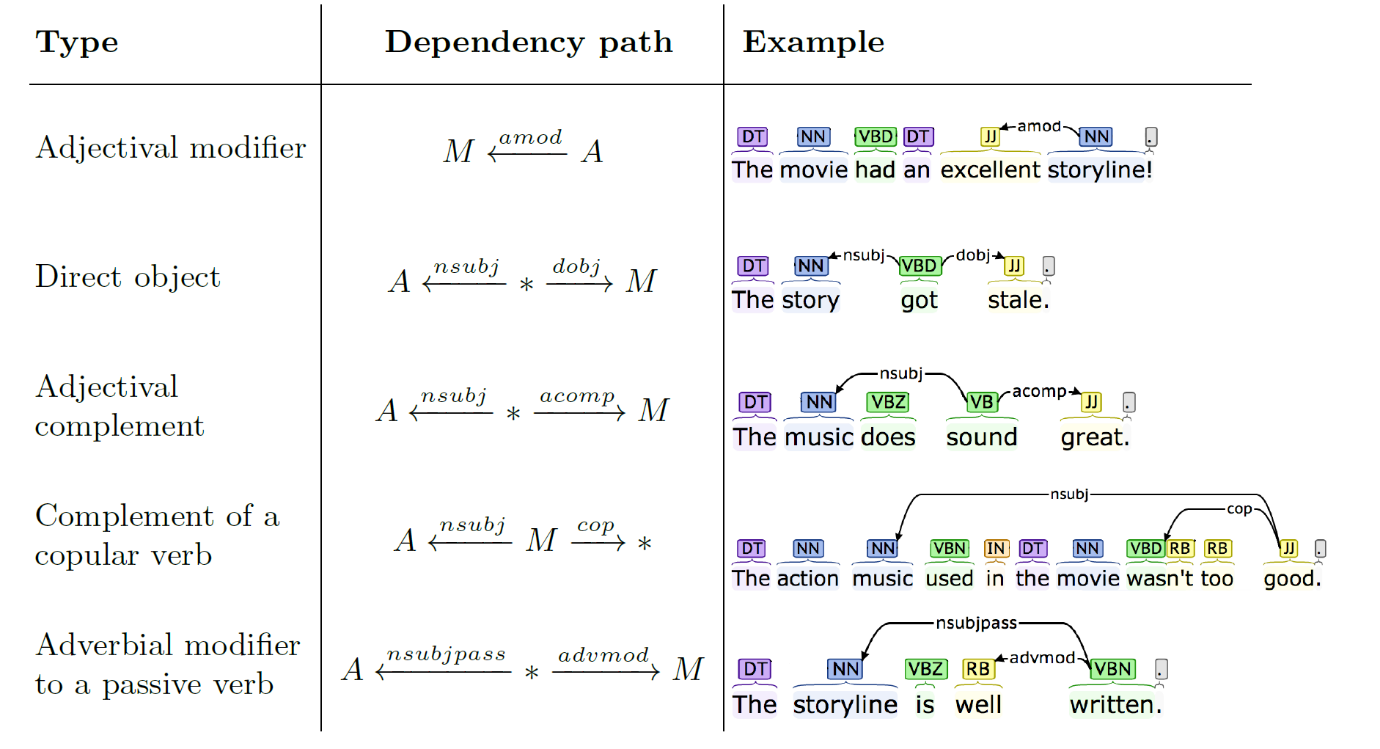
Dalam pengantar di atas telah dijelaskan bahwa metode yang digunakan dalam penelitian ini adalah metode *pattern rule based*. Metode ini memanfaatkan hubungan dari tiap-tiap kata untuk menemukan pasangan aspek dak opini. Hubungan dari tiap kata atau yang lebih sering disebut *dependency rule,* merupakan fitur utama dalam metode ini. *Dependency rule* bisa didapatkan dengan menggunakan library yang sudah ada, sehingga yang dibutuhkan hanyalah text input dan sistem akan otimatis mengeluarkan *dependency rule* dari tiap-tiap kata. *Rule* yang digunakan dalam *study case* ini menggunakan *rule* yang telah didefinisikan oleh Wouter Becken d.k.k didalam papernya yang berjudul Automatically Detecting and Rating Product Aspects from Textual Customer Reviews [1]. Wouter Becken di dalam papernya menggunakan 10 aturan untuk menentukan pasangan aspek dan opini seperti dalam gambar di bawah ini :

Figure 1

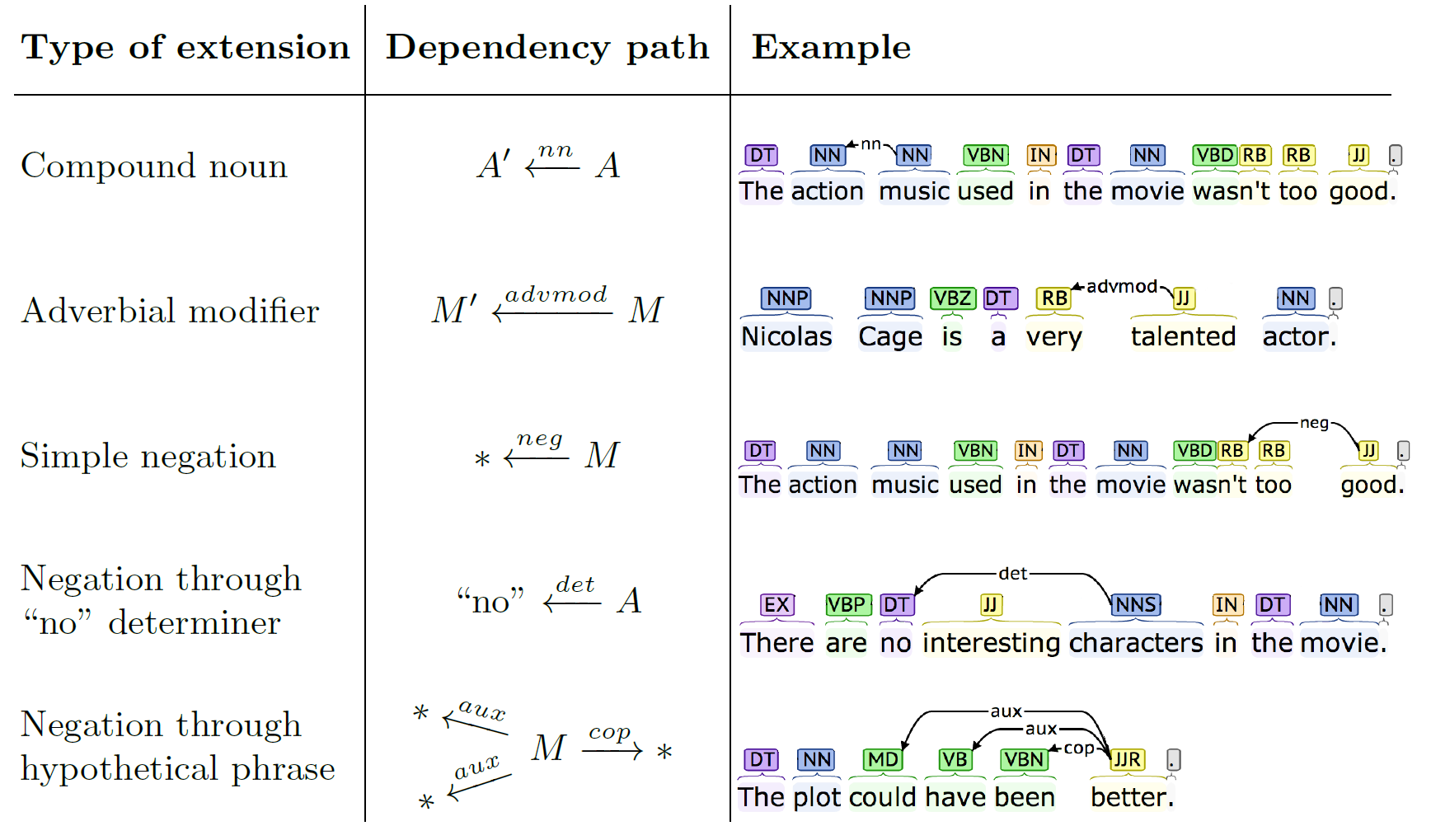


Figure 2

Dimana di dalam gambar pertama terdapat 5 *rule* utama dalam menentukan pasangan aspek dan opini, dan gambar kedua merupakan 5 *rule* tambahan untuk mendapatkan frasa.

## **Implementasi**

Tahapan implementasi secara garis besar di muat dalam flowchart dibawah,

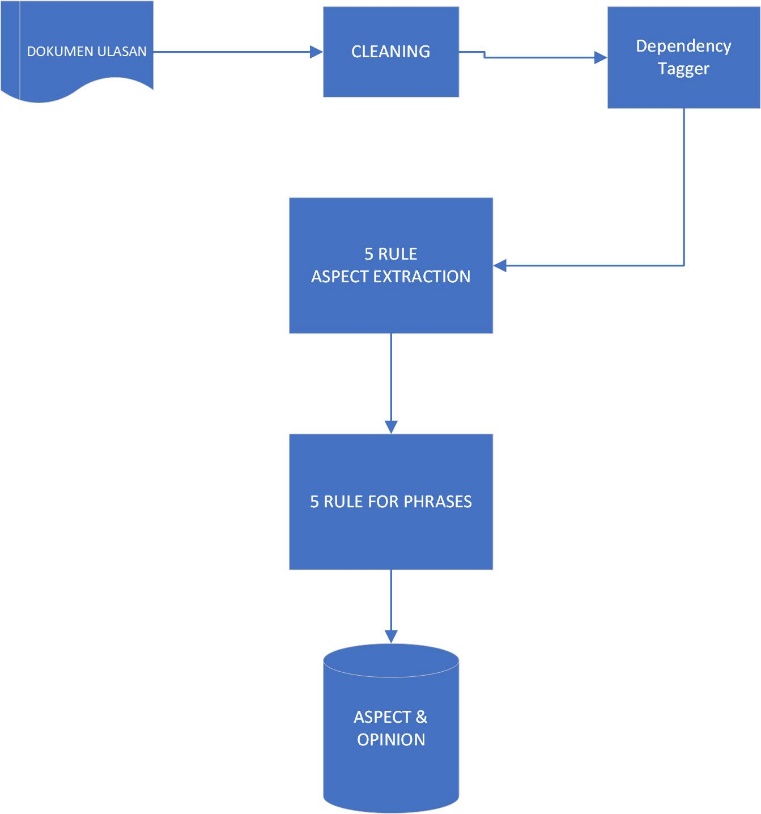


Figure 3

Implementasi dilakukan dengan menggunakan bahasa pemrograman python dengan menggunakan beberapa library pemrosesan bahasa alami seperti nltk dan *stanford dependency parser*.Tahapan yang pertama yang dilakukan adalah proses membersihkan data ulasan, yang dimana data ulasam masih memiliki beberapa simbol dari hasil anotasi seperti ## yang sebelumnya di ikuti oleh aspek yang melekat dalam datanya seperti dalam contoh : “canon powershot g3[+3]##i recently purchased the canon powershot g3 and am extremely satisfied with the purchase “ , dimana dalam proses ini simbol anotasi akan dihilangkan dan aspek yang melekat akan disimpan untuk proses validasi/akurasi. Proses ini dilakukan dalam file clean.py. Output yang dihasilan dalam proses ini adalah sebuah dataset yang bersih berisi 600 baris yang setiap barisnya berisi ulasan yang siap untuk di ekstrak aspek dan opininya.

Tahapan kedua adalah proses *tagger*. Di mana proses bertujuan untuk melabeli setiap kata dengan kata lain berdasarkan hubungan yang dibentuk dari tiap kata dalam kalimat. Seperti dalam gambar 1, kata Action Music memliki relasi NN, relasi dan hubungan kata ini merupakan tujuan utama yang akan dicari dalam proses ini. Proses ini menggunakan Dependency Tag dari library *Stanford Dependency Parser.* Semua keterkatian antar kata akan disimpan untuk proses selanjutnya.

Tahapan ketiga adalah proses implementasi *rule* yang telah didefinisikan dalam paper. 5 Rule pertama bertujuan untuk mencari pasangan aspek dan opini sedangkan 5 Rule berikutnya untuk mencari frasa dalam sebuah aspek maupun opini. Dalam tahapan ini setiap ulasan akan dicek pada setiap 5 rule aspek dan opini, jika kata ulasan masuk kedalam rule tersebut maka tahap selanjutnya memeriksa apakah aspek atau opini tersebut memiliki frasa, sehingga setiap aspek maupun opini bisa lebih dari 1 kata. Setelah semuanya didapat, hasil disimpan dalam list yang berisi pasangan aspek dan opini, di mana hasil ini akan disimpan dalam file .csv untuk proses perhitungan akurasi.

## **Evaluasi dan Akurasi**

1. Menggunakan Data Train

Hasil yang didapat dalam percobaan ini disimpan dalam file Aspect Extraction.csv, di mana hasil yang didapat terlihat tidak begitu meyakinkan, dan hanya beberapa ulasan yang dapat ditemukan pasangan aspek dan opini yang sesuai berdasarkan data yang sudah di anotasi. Perhitungan akurasi berdasarkan presicion dan recall dari tiap-tiap data ulasan. Berikut hasil visualisasi yang didapat dari 600 Ulasan

Figure 4

Dari Plot diatas dapat dilihat sekilas tentang hasil presisi yang diperoleh dari rentang 0-1, dimana terdapat beberapa ulasan yang berhasil di ekstrak sempurna, namun jika terdapat banyak ulasan yang gagal ditemukan pasangan aspek dan opininya, dimana rata-rata dari *presicion* adalah 0.22,

Sedangkan untuk recall dapat di visualisasikan dalam gambar di bawah,

Figure 5

Tidak jauh berbeda dengan *presicion*, recall juga memiliki akurasi yang sedikit yaitu 0.21. Dari 2 hasil diatas dapat disimpulkan bahwa sistem belum bekerja dengan maksimal yang di mana hasil, hasil pembahasan dan analisis akan dibahas di bab berikutnya.

1. Menggunakan Data Test

Pengujian yang kedua menggunakan data test, berisi 14 kalimat dengan fitur yang sudah dianotasikan didalam kalimat tersebut. Pengujian menggunakan data test dihasilkan bahwa 11 dari ke 14 kalimat berhasil diekstrak fiturnya, namun terdapat kalimat yang tidak dapat diketahui fiturnya dan terdapat kegagalan pada kalimat tertentu yang meng-ekstrak fitur lebih dari 1 sehingga dari ke 11 tersebut masih terdapat kegagalan, dengan nilai rata-rata recall sebesar 78% dan presisi sebesar 64%. Nilai yang terbilang cukup tinggi jika dibandingkan dengan hasil pengujian dengan data test. Hal ini diakibatkan pola kalimat dari datatest yang cenderung lebih sederhana dan tidak terlalu kompleks, kalimat yang sederhana cenderung dengan mudah akan diekstrak oleh sistem, sehingga dapat meningkatkan akurasi dari sistem yang dibangun.

## **Analisis dan Kesimpulan**

Metode *pattern rule based* memiliki beberapa kelebihan dan kekurangan dalam proses implementasi maupun dari hasil yang diperoleh, salah satu kekurangan yang paling terasa dalam metode ini dalah *noise* yang dihasilkan. Metode ini hanya terbatas oleh aturan yang telah didefinisikan oleh para peneliti yang sudah melakukan proses *Aspect Extraction*. Rule yang dihasilkan oleh para peneliti diperoleh dari hasil pengetahuan mereka terhadap sebuah kalimat dan pola kalimat. Dalam penelitian ini *rule* utama dalam menemukan pasangan aspek dan opini hanya terbatas dalam 5 *rule*. Dimana setiap *rule* ini bersifat “kaku” sehingga jika ada sebuah kalimat dengan konteks yang berbeda dari *rule,* maka kalimat tersebut akan gagal dikenali aspek dan opininya. Terlebih lagi dalam kasus yang digunakan, ulasan kebanyakan lebih dari 1 kalimat sehingga rule yang didefinisikan di awal cenderung mengekstrak beberapa kata yang tidak sesuai, di mana hal ini diakibatkan oleh level kalimat yang bertambah akan menghasilkan lebih banyak variasi tagger, sehingga baik kalimat 1 dan 2 dalam sebuah ulasan keterhubunggannya juga akan sangat kompleks sehingga akan dihasilkan banyak tag, dari tag tersebut belum tentu merupakan pasangan aspek dan opini namun hal ini cenderung akan di ekstrak oleh sistem. Dalam kasus 1 kalimat dengan konteks yang jelas metode ini sangat baik untuk di implementasikan namun, jika ulasan sudah melebihi dari 2 kalimat dan ada kalimat diluar dari *rule* yang sudah didefinisikan maka kemungkinan gagal akan semankin tinggi. Hal inilah yang mengakibatkan rendahnya akurasi yang diperoleh.

Berdasarkan hasil akurasi dari impelentasi yang sudah dilakukan terdapat beberapa hal yang diapat disimpulkan :

1. Metode *rule base* hanya terbatas dalam aturan saja untuk menemukan pasangan aspek dan opini sehingga akan gagal dalam mengenali aspek/opini jika kalimat tersebut berada di luar rule
2. Hasil yang diperoleh tergantung dari komplesitas kalimat, semakin kompleks kalimat yang akan di uji maka sistem akan semakin sulit mengenali pasangan aspek dan opini yang sesuai.
3. Hasil yang diperoleh dalam kasus ini terbilang rendah dikarenakan banyak kalimat dengan komplesitas yang tinggi, dan konteks kalimat berada di luar *rule* yang telah didefinisikan.

## **Daftar Pustaka**

[1] Bancken,Alfarone,Davis. (2014). Automatically Detecting and Rating Product Aspect from Textual Costumer Review. Department of Computer Science, KU Leuven Celestijnenlaan 200A