INTISARI

DETEKSI *TRENDING TOPIC TWEET* BERBAHASA INDONESIA MENGGUNAKAN METODE *CLUSTERING* SERTA KOMBINASI *TEXTUAL* DAN *SOCIAL CONTENT*

Oleh

Indra
13/351298/SPA/00466

Deteksi *trending topic* menggunakan tiga pendekatan yaitu berbasis *textual content*, *social content* dan *hybrid*. Ketiga metode tersebut memiliki permasalahan yang berbeda. Pertama, deteksi *trending topic* berbasis *textual content* memiliki permasalahan dalam menggunakan prapemrosesan yang kompleks. Kedua, deteksi *trending topic* berbasis *social content* belum mampu mendeteksi konten *trending topic*. Ketiga, deteksi *trending topic* berbasis *hybrid* sangat dipengaruhi oleh pengguna Twitter dengan jumlah *follower* yang besar (jumlah *follower* ribuan bahkan jutaan).

Pada penelitian ini dilakukan penerapan metode *baseline* berbasis *textual content* yaitu BN-grams dan Doc-p pada *tweet* berbahasa Indonesia. Selanjutnya, BN-grams dilakukan modifikasi pada pembentukan klaster dan perangkingan topik menjadi metode *Non Overlap Trending Topic* (NOTT) dan *Overlap Trending Topic* (OTT). Kemudian, dilakukan penggabungan antara *Link Anomaly* (berbasis *social content*) dan deteksi *burst Kleinberg* (berbasis *textual content*) dengan hasil akhir berisi *intersection* interval waktu antara *Link Anomaly* dan deteksi *burst Kleinberg* serta menjadi metode baru yaitu *Overlap Time Interval Trending Topic* (OTITT).

Metode NOTT dan OTT memiliki empat tahapan yang sama yaitu: prapemrosesan, pembentukan klaster menggunakan *Frequent Term Based Clustering* (FTC) atau *Hierarchical Frequent Term Based Clustering* (HFTC), perangkingan topik dan pemodelan topik. Metode OTITT terdiri dari empat tahapan utama yaitu: ekstraksi interval waktu dari *Link Anomaly*, ekstraksi interval waktu dari deteksi *burst Kleinberg*, *intersection* interval waktu *Link Anomaly* dan deteksi *burst Kleinberg* serta pemodelan topik. Keseluruhan metode usulan dilakukan pengujian dengan membandingkan hasil *trending topic* metode usulan dengan *trending topic* yang berasal dari media siber maupun Twitter.

Berdasarkan hasil pengujian, metode BN-grams memiliki nilai topic recall lebih tinggi dibandingkan Doc-p dengan nilai 55%. Disisi lain, metode OTT memiliki nilai topic recall lebih tinggi dibandingkan NOTT, BN-grams dan Doc-p dengan nilai 33%. Metode OTITT memiliki nilai topic recall 47% dan tertinggi dibandingkan NOTT, OTT, BN-grams maupun Doc-p.

Kata kunci: OTITT, BN-grams, *Link Anomaly*, NOTT, OTT.

ABSTRACT

TRENDING TOPIC DETECTION OF INDONESIAN TWEET USING CLUSTERING METHOD AND COMBINATION OF TEXTUAL AND SOCIAL CONTENT

By

Indra

13/351298/SPA/00466

Detection of trending topic using three approaches that are textual content, social content, and *hybrid*. The three methods have different problems. First, trending topic detection based on textual content has difficulties in using complex preprocesses. Secondly, trending topic detection based on social content has not been able to detect trending topic content. Thirdly, trending topic detection with the *hybrid* based method is influenced by Twitter *user*s with considerable influence (number of followers thousands and even millions).

In this research, we explore of textual content baseline method with BN-grams and Doc-p on Indonesian *tweet*s. Furthermore, we modify BN-grams in the step of generating a cluster and topics ranking into Non-Overlap Trending Topic (NOTT) and Overlap Trending Topic (OTT) methods. Then, we combine between Link Anomaly (based on social content) and Burst Kleinberg detection (based on textual content) with the final result containing the intersection of time intervals between Link Anomaly and detection of Kleinberg bursts and becoming a new method of Overlap Time Interval Trending Topic (OTITT).

The NOTT and OTT method has the same four steps, i.e., preprocessing, cluster formation using Frequent Term Based Clustering (FTC), Hierarchical Frequent Term Based Clustering (HFTC), topic ranking and topic modeling. The OTITT method consists of four main stages: extraction of time intervals from Link Anomaly, extraction of time intervals from detection of burst Kleinberg, the intersection of Link Anomaly time intervals and detection of Kleinberg bursts and topic modeling. We evaluate the entire proposed method by comparing the results of the proposed method trending topic with trending topics originating from cyber media and Twitter.

Based on the experimental results, BN-grams method has a higher topic recall value than Doc-p with 55%. On the other hand, the OTT method has more topic recall value than NOTT, BN-grams, and Doc-p with 33%. The OTITT method has a 47% topic recall topic and is highest than NOTT, OTT, BN-grams or Doc-p.

Keywords: OTITT, BN-grams, Link Anomaly, NOTT, OTT.